武漢大學
WUHAN UNIVERSITY

武汉大学

本科生课程讲义及知识总结

课程名称： 空间数据分析
开课学院： 遥感信息工程学院
开课时间： 2020-2021 年度第二学期

2021. 武汉大学

目录

第一章·绪论

➢ 大数据、科学研究的范式与空间分析

1. **大数据及其发展趋势**:

   定义：大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合

   发展趋势：大数据产业正在成为新的经济增长点。当前，在大数据应用的实践中，描述性、预测性分析应用多，决策指导性等更深层次分析应用偏少；未来，随着应用领域的拓展、技术的提升、数据共享开放机制的完善，以及产业生态的成熟，具有更大潜在价值的预测性和指导性应用将是发展的重点。

2. **科学研究的四种范式**:

   1) **实验型(experimental)范式**：亚里斯多德以来的上千年历史；
   2) **理论型(theoretical)范式**：牛顿依赖的几百年历史；
   3) **计算机仿真型(computational)范式**：计算机发明依赖的几十年历史；
   4) 第四范式：数据密集型的科学发现(data-intensive scientific discovery)

   - **第三范式与第四范式的显著区别在于：**计算科学是先提出可能的理论，再搜集数据，然后通过计算仿真进行理论验证；而数据密集型科学，是先有了大量的已知数据，然后通过计算得出之前未知的理论

3. **空间分析的定义（郭仁忠院士，1997)**：空间分析是基于地理对象的位置和形态特征的空间数据分析技术，其目的在于提取和传输空间信息。

4. **GIS 被定义为**：地理信息系统是一种特定而又十分重要的空间信息系统，它是以采集、存储、管理、分析和描述整个或部分地球表面(包括大气层在内)与空间和地理分布有关的数据的计算机空间信息系统。

5. **空间分析是 GIS 的核心和灵魂：**①空间分析是地理信息系统的主要特征，是区别地理数据库和地理信息系统的标准，是评价一个地理信息系统的主要指标之一；②空间分析是 GIS 的核心。空间数据的采集、存储和管理是为空间分析提供数据基础，空间数据的描述是空间分析结果的表达，是对经过数据预处理的空间数据的深层次分析和处理；③GIS 的发展需要理论和技术共同发展，空间分析兼具理论性和技术性，是 GIS 发展的重要突破口；④GIS 已经从数据库型 GIS 发展成分析型 GIS 阶段，空间分析功能成为人们关注的焦点。

第二章·空间分析理论

➢ 地理学定律

1. **地理学第一定律（空间相关性）**：任何事物都是空间相关的，距离近的事物比距离远的事物的空间相关性更大。（Tobler，1970）
2. **地理学第二定律（空间异质性）**：地理现象具有不可控的空间变化。（Michael F. Goodchild, 2004）
3. **地理学第三定律（空间相似性）**：地理环境越相似，地理目标特征越接近。（A-xing Zhu et al., 2018）
4. **空间邻近度**：空间邻近度正比于公共边界长，反比于中心距。
5. **时空邻近度**：地理空间任意两匀质区域（含点）之间的时空邻近度，对给定的“流”，正比于二者之间的总流量，反比于从一端到达另一端的平均时间。

➢ 空间关系理论

1. **空间关系**：空间关系是指空间对象之间的各种几何关系。
2. **空间关系的类型**：①顺序空间关系（描述目标在空间中的某种排序，如东西南北等）；②度量空间关系（用某种度量空间中的度量来描述的目标间的关系，如距离关系）；③拓扑空间关系（拓扑变换下的拓扑不变量，如空间目标的相邻和连通关系）。
3. **空间关系的约束强度**：度量关系>顺序关系>拓扑关系
4. **度量空间关系**：定性化描述和定量化测度两种
   1) **定性化描述**：远、近、很远
   2) **定量化测度**：①空间指标量算：用区域空间指标量测空间目标间的空间关系（几何指标、自然地理参数、人文地理指标）；②距离度量：利用距离量算目标间的空间关系（可以沿着实际地球表面也可以沿着地球椭球体的距离量算）。
5. **拓扑变换（又可比喻为橡皮几何学）**：在图形被弯曲、拉大、缩小或任意变形下，图形原来的点与变换后的点一一对应，并且邻近点还是邻近点。这样的变换称作拓扑变换。【两个条件：①变化前后点一一对应；②不产生新点也不减少点】
6. **空间关系描述**：以数学或逻辑的方法区分不同的空间关系，给出形式化的描述。
7. **顺序空间关系描述：**
   1) **两点之间的基本方向关系**：

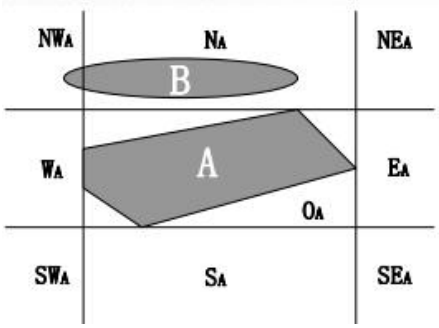
例如正东可以描述为：$restricted_{east(p_i,q_i)} \equiv X(p_i) > X(q_j) \wedge Y(p_i) = Y(q_j)$

2) **方向关系的定性描述模型**：①锥形模型；②最小约束矩形模型（MBR，Minimum Bounding Rectangle）；③方向关系矩阵模型

✧ **锥形模型**：在从某个空间目标出发指向另一个目标的锥形区域中确定两个空间目标间的空间方向关系。（从参考目标的质心出发作两条相互垂直的直线，将所在的平面划分为 4 个无限锥形区域，每个锥形顶点的角平分线指向方向为一个主方向(东南西北)）

✧ **最小约束矩形模型**：利用两个目标间的 MBR 间的关系定义方向关系，利用空间对象的几何近似关系取代实际空间对象的关系。

✧ **方向关系矩阵模型**：将平面空间划分为 9 个区域，每个区域为一个方向片，每个方向片对应一个主方向，参考目标所在的方向片称为同方向。物体 A 的方向集：$\{NW_A, N_A, NE_A, W_A, O_A, E_A, SW_A, S_A, SE_A\}$。将 B 与 A 的九个方向片分别求交，得到方向关系矩阵。



$$Dir(A,B) = \begin{bmatrix} NW_A \cap B & N_A \cap B & NE_A \cap B \\ W_A \cap B & O_A \cap B & E_A \cap B \\ SW_A \cap B & S_A \cap B & SE_A \cap B \end{bmatrix}$$

$$\begin{bmatrix} \frac{Area(NW_A \cap B)}{Area(B)} & \frac{Area(N_A \cap B)}{Area(B)} & \frac{Area(NE_A \cap B)}{Area(B)} \\ \frac{Area(W_A \cap B)}{Area(B)} & \frac{Area(O_A \cap B)}{Area(B)} & \frac{Area(E_A \cap B)}{Area(B)} \\ \frac{Area(SW_A \cap B)}{Area(B)} & \frac{Area(S_A \cap B)}{Area(B)} & \frac{Area(SE_A \cap B)}{Area(B)} \end{bmatrix}$$

用面积比代替交集

8. **度量空间关系描述**：

1) **空间指标量算**：包括长度、周长、面积等指标。

2) **距离度量描述**：以两个点目标间的距离为度量。

a) **平面中两点距离**：欧式距离、契比雪夫距离、马氏距离、明氏距离。

✧ **欧氏距离**：$d(A,B) = \|A - B\| = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$

✧ **契比雪夫距离**：$d(A,B) = \max_i |a_i - b_i|$

✧ **马氏距离**：$d(A,B) = |x_1 - x_2| + |y_1 - y_2|$

（**曼哈顿距离**：纬度差+经度差）

✧ **明氏距离**：$d(A,B) = (|x_1 - x_2|^m + |y_1 - y_2|^m)^{1/m}$

b) **球面距离**：大地测量距离（球面上两点间的大圆距离；大圆：地球表面二点

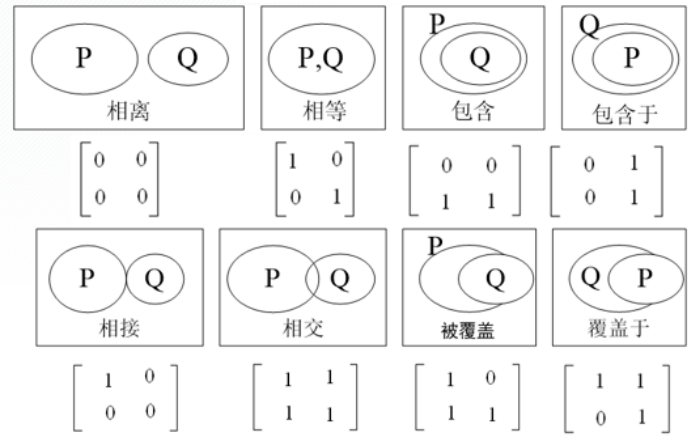
与球心构成的平面构成的大圆圈）

c) **具有行业特色的距离定义**：①统计学中的斜交距离、马氏距离等；②旅游业中的旅游时间距离（两个点(如两个城市)之间的旅游时间距离为从一个点(城市)到另一个点(城市)的最短时间）。

9. **拓扑空间关系描述**：①4 元组模型；②9 元组模型；③V9I 模型

1) **4 元组模型**：将简单空间实体看作是边界点和内部点构成的集合，4 元组模型为由两个简单空间实体点集的边界与边界的交集、边界与内部的交集、内部与边界的交集、内部与内部的交集构成的 4 元组（2×2 矩阵，边界用∂A，内部用A°）

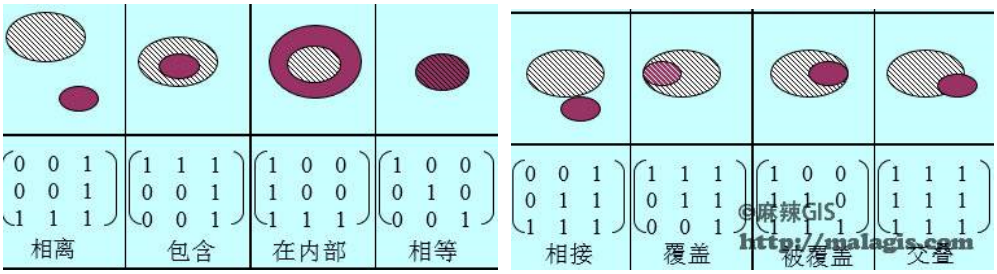
$$R(A,B)=\begin{bmatrix}\partial A\cap\partial B & \partial A\cap B^{\circ}\\ A^{\circ}\cap\partial B & A^{\circ}\cap B^{\circ}\end{bmatrix}$$



2) **9 元组模型**：在 4 元组的基础上，在空间描述框架中引入空间实体的“补（外部，用A^{-}）”的概念，将空间目标 A 表示为边界、内部和外部三个部分的集合。

$$R(A,B)=\begin{bmatrix}\partial A\cap\partial B & \partial A\cap B^{\circ} & \partial A\cap B^{-}\\ A^{\circ}\cap\partial B & A^{\circ}\cap B^{\circ} & A^{\circ}\cap B^{-}\\ A^{-}\cap\partial B & A^{-}\cap B^{\circ} & A^{-}\cap B^{-}\end{bmatrix}$$

9 元组模型描述 8 种面面间关系：



3) **V9I 模型**：用 Voronoi 多边形取代 9 元组中的“补”重新定义 9 元组模型，并将其定义为 V9I 模型。（空间目标O_i的 Voronoi 区域可以定义为：到目标O_i的距离比到所有其它目标的距离都近的点所构成的区域）

$$R(A,B)=\begin{bmatrix}\partial A\cap\partial B & \partial A\cap B^{\circ} & \partial A\cap B^{V}\\ A^{\circ}\cap\partial B & A^{\circ}\cap B^{\circ} & A^{\circ}\cap B^{V}\\ A^{V}\cap\partial B & A^{V}\cap B^{\circ} & A^{V}\cap B^{V}\end{bmatrix}$$

10. **时空空间关系：**地理实体之间的空间关系往往随着时间而变化，时间关系交织在一起就形成了多种时空关系。

    时空区间逻辑（13 种）：时间相等 TR_equal、时间前 TR_before、时间后 TR_after、时间相遇 TR_meet、时间被遇见 TR_met、时间交叠 TR_overlap、时间被交叠 TR_overlapped、时间包含 TR_contain、时间被包含 TR_during、时间开始 TR_start、时间被开始 TR_started、时间终止 TR_finish、时间被终止 TR_finished。

➢ 空间认知理论

1. **地理空间认知**：是指在日常生活中，人类如何逐步理解地理空间，进行地理分析和决策，包括地理信息的知觉、编码、储存以及解码等一系列心理过程。
2. **地理知觉：**指将地理事物从地理空间中区分出来，获取其位置并对其进行识别。

   格式塔心理学（Gestalt Psychology）：现代认知心理学的先祖，又称完形心理学，是一种研究经验现象中的形式与关系的心理学，强调整体结构感知和自顶向下的加工。
3. **地理空间知觉方法**：环境空间的知觉主要靠导航经验，地理空间的知觉主要靠读图。

   ① 基于地图的地理空间认知：通过阅读地图实现人对地理空间的认知；

   ② 导航方式：通过导航经验感觉环境空间。【二者存在较大差异】
4. **表象 vs 地理表象**：
   1) **表象**：认知科学的重要概念，是人类意识对物质世界主动和积极的形象化反映。
   2) **地理表象**：表示地理形象思维所产生的各种“象”，是地理思维活动的产物，地理思维得以进行的载体。
5. **概念化 vs 地理概念化**：
   1) 概念化：把具有共同特征的事物归为一类，而把不同特征的事物放在不同类中。
   2) 地理概念化：是地理世界已知地理实体、实体属性和实体间关系的知识库，依据概念化知识记忆和理解地理世界。
6. **地理概念化的方法**：
   1) **基于经典集合论的地理概念化方法**：概念形式化定义为一个三元组(O,A,R)
   2) **基于原型的地理概念化方法**：关于某一类事物的典型特征模式，物体特征与原型认知范畴越接近，就越有可能被划归到某一原型范畴中。

➢ 空间推理理论与空间不确定性理论

1. **空间推理：**指利用空间理论和人工智能技术对空间对象进行建模、描述和表示，并据此对空间对象间的空间关系进行定性或定量分析和处理的过程。
2. **地理空间推理的主要研究内容**：
   1) 根据空间目标的位置，基于给定的空间关系形式化表示模型，推断空间目标之间的空间关系；
   2) 根据空间目标之间的已知基本空间关系，推断空间目标之间未知的空间关系；
   3) 利用空间推理从空间数据库中挖掘空间知识，也可以利用事件推理（案例推理）的方法进行空间目标的模糊查询；
   4) 基于常识的空间推理以及加入时间因素的时空推理。
3. **空间推理的方法**：①不确定性推理；②概率推理；③贝叶斯推理；④可信度推理；⑤证据推理；⑥模糊推理；⑦案例推理；⑧空间关系推理；⑨时空推理。
4. **贝叶斯推理：**已知证据 E 的概率$P(E)$和假说 H 的先验概率$P(H)$，并已知 H 成立时 E 出现的条件概率$P(E|H)$ ➔ 从 H 的先验概率 P(H)推得 H 的后验概率：

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

➔ 如果一个证据 E 支持多个假设 H_1 H_2 H_3 … H_i：

$$P(H_i|E) = \frac{P(H_i) \times P(E|H_i)}{\sum_{j=1}^{n}\left(P(H_j) \times P(E|H_j)\right)}$$

贝叶斯推理的计算案例：

某地区居民的肝癌发病率为 0.0004，现用甲胎蛋白法进行普查。已知患有肝癌的人其化验结果 99%呈阳性(有病)，而没患肝癌的人其化验结果 99.9%呈阴性(无病) ➔ **在化验结果呈阳性的人中可能有多少人患有肝癌?**

➔ E 表示样本的观察证据“化验结果呈阳性”；H_1表示假说命题“被检查者患有肝癌”，H_2表示假说命题“被检查者没有患肝癌”。

➔ $P(H_1)$（即某地区居民的肝癌发病率)=0.0004

$P(H_2)$（即某地区居民没患肝癌的比率)=1-0.0004=0.9996

$P(E|H_1)$（即患有肝癌者其化验结果呈阳性的比率)=0.99

$P(E|H_2)$（即没患肝癌者其化验结果呈阳性的比率)=1-0.999=0.001

➔ 推断$P(H_1|E)$：化验结果呈阳性的条件下，假说“被检查者患有肝癌”的比率。

$$P(H_1|E)=\frac{P(H_1)P(E|H_1)}{P(H_1)P(E|H_1)+P(H_2)P(E|H_2)}=\frac{0.0004\times 0.99}{0.0004\times 0.99+0.9996\times 0.001}=0.284$$

即：在化验结果呈阳性的人中，真患肝癌的人不到 30%

5. **不确定性**：指被测量对象知识缺乏的程度，通常表现为随机性和模糊性。
   1) 随机性：可重复观察，在观察之前知道所有可能的结果，但不知道到底哪一种结果会出现。
   2) 模糊性：两个有区别的概念之间的区别是渐变的而不是突变的，两者之间并不存在明确的界限。
6. 不确定性是地球空间信息基础理论的主要组成部分之一，空间数据的不确定性分析是地球空间信息科学的重要基础理论之一。空间数据的获取和处理产生不确定性，空间数据及分析中的不确定性直接影响到 GIS 产品的质量。

第三章 · 栅格分析与图像挖掘

➢ 栅格数据分析

1. **栅格数据分析**：基于栅格数据的空间分析。

   栅格数据分析的数学基础：线性代数的二维数字矩阵分析方法。

   栅格数据分析的基础理论：地图代数。
2. **栅格数据分析的主要方法**：①聚类分析、②聚合分析、③信息复合分析、④窗口分析、⑤量算分析。
3. **栅格数据的聚类分析**：根据设定的聚类条件对原有数据系统进行有选择的信息提取而建立新的栅格数据系统的方法。可以对单一层面的栅格数据进行聚类分析，也可以对多个层面的栅格数据进行聚类分析。
4. **栅格数据的聚合分析**：根据空间分辨率和分类表，进行数据类型的合并或转换以实现空间地域的兼并。【将较复杂的类别转换为较简单的类别】
5. **栅格数据的信息复合分析**：进行同地区多层面空间信息的自动复合叠置分析。

   2 种信息复合模型：①视觉信息复合；②叠加分类模型

   视觉信息复合：将不同专题的内容叠加显示在结果图件上，以便系统使用者判断不同专题地理实体的相互空间关系。例如遥感影像与专题地图的复合。

   ✧ 视觉信息的叠加不产生新的数据层面，只是将多层信息复合显示，便于分析。

   叠加分类模型：根据参加复合的数据平面各类别的空间关系重新划分空间区域。

6. **栅格数据的窗口分析**：对栅格数据系统开辟一个分析窗口，在该窗口内进行极值、均值等统计计算，或与其它层面的信息进行复合分析。【最大值、最小值、、均值、中值、范围、总和、方差、频数、众数等】
7. **栅格数据的量算分析**：基于遥感图像数据(栅格)可以计算某种地物类型(如耕地)所占的面积或对于栅格格式的 DEM 数据，可以方便进行体积计算等。
8. **栅格数据分析的代表性应用：**①物迁徙路径分析；②夜光遥感图像分析

➢ 图像数据挖掘

1. **图像（遥感图像）数据挖掘与知识发现**：指利用空间数据挖掘的理论和方法（空间聚类分析、空间关联规则分析、空间序列分析等）从图像或图像数据库中提取出规律性的潜在有用的信息、图像数据关系、空间模式等，挖掘出规律性知识，从而为图像的智能化处理和应用决策服务的过程。
2. **概念**：概念是反映对象本质属性的思维形式。

   把所感知的事物的共同本质特点抽象出来，加以概括就成为概念。
3. **图像数据挖掘的主要方法**：①统计/空间统计方法；②归纳方法；③空间聚类方法；④关联规则挖掘方法；⑤探索性数据分析方法；⑥粗糙集方法；⑦云模型方法；⑧概念格方法；⑨决策树方法；⑩神经网络/深度学习方法……
4. **关联规则挖掘方法**：

   1) **支持度 s：**在事务集 D 中同时包含 A 和 B 的事务的百分比

   $$support(A \Longrightarrow B) = P(A \cup B)$$

   2) **置信度 c：**在事务集 D 中包含 A 的事务同时也包含 B 的百分比

   $$confidence(A \Longrightarrow B) = P(B|A)$$

   例如：假设有 6 位顾客在某一零售店内一共购买了 6 种商品。这 6 位顾客记作 T1, T2, T3, T4，T5, T6，6 种商品为：牙刷、筷子、毛巾、香皂、牙膏、杯子，分别记为：a, b, c, d, e, f，可得到如图所示的事务数据列表

TID	项集列表
T1	a，b，d
T2	a，b，c，d
T3	a，b，d，e
T4	b，e，f
T5	a，b，d，f
T6	a，c，d，e

规则	支持度	置信度	规则	支持度	置信度	规则	支持度	置信度
$a \Rightarrow b$	0.67	0.8	$b \Rightarrow a$	0.67	0.8	$c \Rightarrow a$	0.33	1.0
$a \Rightarrow d$	0.83	1.0	$d \Rightarrow a$	0.83	1.0	$a \Rightarrow bd$	0.67	0.8
$ab \Rightarrow d$	0.67	1.0	$ac \Rightarrow d$	0.33	1.0	$b \Rightarrow d$	0.67	0.8
$d \Rightarrow b$	0.67	0.8	$b \Rightarrow ad$	0.67	0.8	$ad \Rightarrow b$	0.67	0.8
$c \Rightarrow ad$	0.33	1.0	$ae \Rightarrow d$	0.33	1.0	$f \Rightarrow b$	0.33	1.0
$c \Rightarrow d$	0.33	1.0	$d \Rightarrow ab$	0.67	0.8	$bd \Rightarrow a$	0.67	1.0
$cd \Rightarrow a$	0.33	1.0	$de \Rightarrow a$	0.33	1.0			

关联规则：$a \Longrightarrow b, s = 0.67, c = 0.8$ ➔ 表示同时购买“牙刷、筷子”的顾客占67%，购买“牙刷”的顾客同时还购买“筷子”的可能性占80%

5. **Apriori 算法：**一种使用频繁项集的先验知识从而生成关联规则的一种算法，也是最有影响的关联规则挖掘算法。

   1) **项集：**包含 k 个项的项集称为 k-项集，所谓项集就是项的集合。

   2) **频繁项集：**如果事务集 D 中项集出现的频率不小于最小支持度阈值与 D 中事务总数的乘积（例如阈值为 0.3，事务总数为 6，则阈值为$0.3 \times 6 = 1.8$，取整为 2），则称它为频繁项集。

   3) **Apriori 算法的基本计算过程**：

      a) 计算所有的C_1（候选 1-项集）

      b) 扫描数据库，根据最小支持度阈值与 D 中事务总数的乘积删除其中的非频繁子集，生成L_1（1-频繁项集）

      c) 将L_1与L_1链接生成C_2（候选 2-项集）

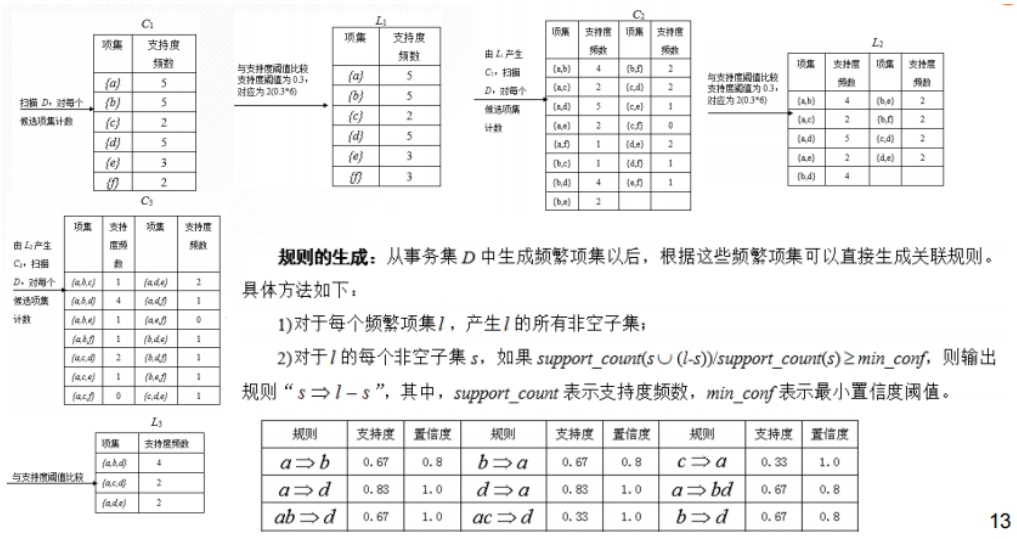
      d) 扫描数据库，删除其中的非频繁子集，生成L_2（2-频繁项集）

      e) 以此类推直至生成L_k（k-频繁项集），此时不再有频繁项集产生

      f) 根据频繁项集生成关联规则：①对于每个频繁项集l，产生l的所有非空子集；②对于l的每个非空子集 s，如果支持度频数$support_{count}$和最小置信度阈值min_{conf}满足下列条件，则输出规则：“$s \Longrightarrow l - s$”

$$\frac{support_{count}(s \cup (l - s))}{support_{count}(s)} \geq min_{conf}$$

   4) **Apriori 算法的应用实例：**



C_1

项集	支持度频数
{a}	5
{b}	5
{c}	2
{d}	5
{e}	3
{f}	2

扫描 D，对每个候选项集计数 → 与支持度阈值比较，支持度阈值为 0.3，对应为 2(0.3*6)

L_1

项集	支持度频数
{a}	5
{b}	5
{c}	2
{d}	5
{e}	3
{f}	3

由 L_1 产生 C_2，扫描 D，对每个候选项集计数

C_2

项集	支持度频数	项集	支持度频数
{a,b}	4	{b,f}	2
{a,c}	2	{c,d}	2
{a,d}	5	{c,e}	1
{a,e}	2	{c,f}	0
{a,f}	1	{d,e}	2
{b,c}	1	{d,f}	1
{b,d}	4	{e,f}	1
{b,e}	2		

与支持度阈值比较，支持度阈值为 0.3，对应为 2(0.3*6)

L_2

项集	支持度频数	项集	支持度频数
{a,b}	4	{b,e}	2
{a,c}	2	{b,f}	2
{a,d}	5	{c,d}	2
{a,e}	2	{d,e}	2
{b,d}	4		

由 L_2 产生 C_3，扫描 D，对每个候选项集计数

C_3

项集	支持度频数	项集	支持度频数
{a,b,c}	1	{a,d,e}	2
{a,b,d}	4	{a,d,f}	1
{a,b,e}	1	{a,e,f}	0
{a,b,f}	1	{b,d,e}	1
{a,c,d}	2	{b,d,f}	1
{a,c,e}	1	{b,e,f}	1
{a,c,f}	0	{c,d,e}	1

与支持度阈值比较 →

L_3

项集	支持度频数
{a,b,d}	4
{a,c,d}	2
{a,d,e}	2

规则的生成：从事务集 D 中生成频繁项集以后，根据这些频繁项集可以直接生成关联规则。具体方法如下：

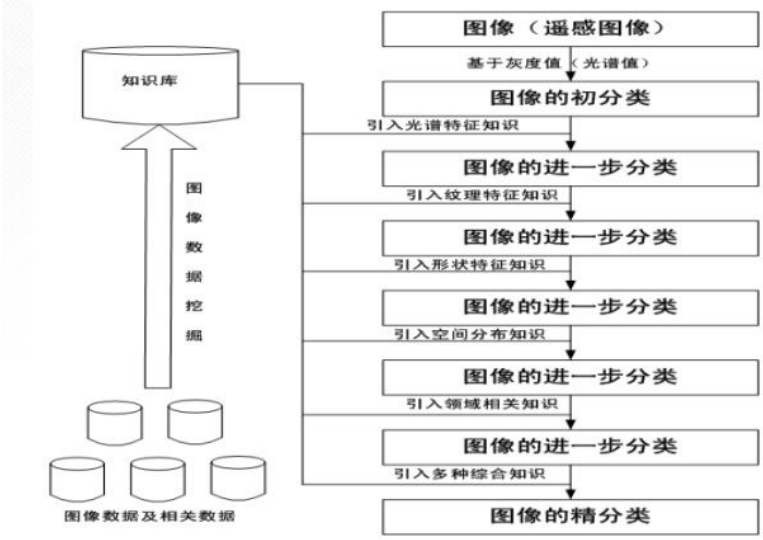
1)对于每个频繁项集 l，产生 l 的所有非空子集；

2)对于 l 的每个非空子集 s，如果 $support\_count(s \cup (l\text{-}s))/support\_count(s) \geq min\_conf$，则输出规则“$s \Rightarrow l - s$”，其中，$support\_count$ 表示支持度频数，min_conf 表示最小置信度阈值。

规则	支持度	置信度	规则	支持度	置信度	规则	支持度	置信度
$a \Rightarrow b$	0.67	0.8	$b \Rightarrow a$	0.67	0.8	$c \Rightarrow a$	0.33	1.0
$a \Rightarrow d$	0.83	1.0	$d \Rightarrow a$	0.83	1.0	$a \Rightarrow bd$	0.67	0.8
$ab \Rightarrow d$	0.67	1.0	$ac \Rightarrow d$	0.33	1.0	$b \Rightarrow d$	0.67	0.8

13

6. **图像挖掘的应用**：①纹理特征关联规则挖掘；②光谱特征关联规则挖掘；③空间分布规律挖掘；④基于知识的图像分类（示意图如下所示）；⑤基于知识的图像检索；⑥基于知识的目标识别……



➢ 夜光遥感分析与挖掘

夜光遥感：在夜间无云情况下，遥感传感器获取陆地/水体可见光源的过程即夜光遥感

2. **夜光遥感的典型应用**：①社会经济参数估算；②城市化和区域发展评估；③夜光遥感评估叙利亚内战；④光污染分析；⑤渔业监测；⑥宗教和文化分析；⑦火点检测；⑧遥感国庆期间人类活动……

3. **分析遥感影像/产品的时空变化一般有两种主流方法**：①不同区域总量在时间维度的变化，但损失空间细节信息；②初始时期和结束时期的影像进行变化检测，损失了时间细节信息。

4. **夜光遥感的分析挖掘方法**：①几何配准；②重投影和夜光总量计算；③城市范围提取

   1) 不同的 DMSP/OLS 稳定夜光影像经常出现几何误差，研究表明误差的可能会在 2-3 个像素左右，通常利用几何配准的方式来消除不同夜光影像之间的几何误差。

   2) 计算区域内夜光亮度的总和，是挖掘夜光遥感经济信息的重要手段。为了计算区域内灯光亮度的总和，需要让不同像素之间的面积相等，常用的手段是利用 Goode Land 投影，或是直接计算每个像素的大小之和。

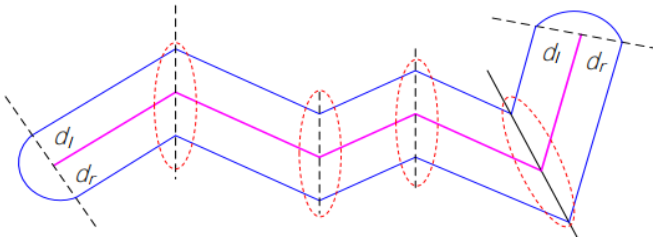
   3) 利用 MODIS、Landsat 等日间遥感影像提取城市范围容易产生同谱异物现象，对夜光遥感影像进行阈值分割则能够快速提取城市范围。

第四章·矢量分析与空间社会网络

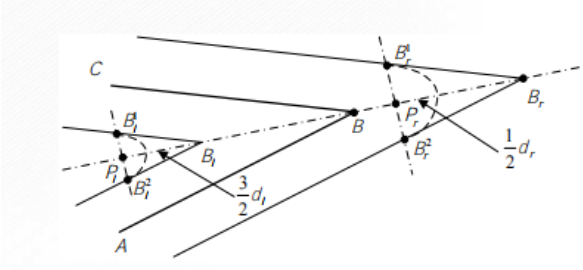
➢ 矢量分析的基本方法

1. **包含分析**：用于确定空间要素之间是否存在直接的空间位置上的联系。
2. **包含分析的类型（6种）**：①点与点的包含关系；②点与线的包含关系；③点与面的包含关系；④线与线的包含关系；⑤线与面的包含关系；⑥面与面的包含关系。
   1) **点与点的包含关系**：计算两点之间的距离；距离为零或小于某个阈值认为包含。
   2) **点与线的包含关系**：一个点落在线状目标上；通过计算点到线之间的距离，距离为零或小于某个阈值认为包含。
   3) **点与面的包含关系**：点完全落在面内；通过判断点是否位于面域范围之内来判断
   4) **线与线的包含关系**：一条线完全或部分包含另一条线。
   5) **线与面的包含关系**：线完全落在面内；通过判断组成该线的所有节点是否都包含在某个面之内。【多个点与面之间的包含关系问题】
   6) **面与面的包含关系**：一个面完全被另一个面包含；通过判断组成一个面的所有节点是否都包含在另外一个面的区域范围之内。【多个点与面之间的包含关系问题】
3. **缓冲区分析**：以空间要素（点、线、面或复杂要素）为基础，基于近邻的概念按照一定的缓冲区距离向外/向内扩展一定的范围并进行分析，位于缓冲区距离之内的区域为缓冲区，之外的区域为非缓冲区。

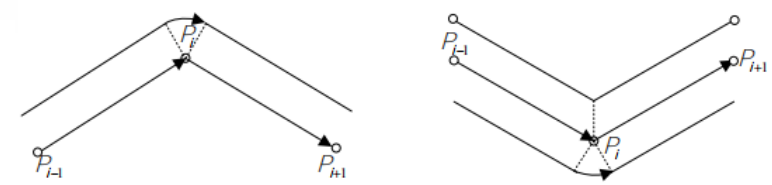
   数学观点：空间目标O_i的缓冲区的数学定义为：$B_i = \{x: d(x, O_i) \leq R\}$
4. **缓冲区分析分两个部分**：①缓冲区域的生成；②在缓冲区内进行统计分析或查询分析
5. **缓冲区的类型**：
   1) **点要素的缓冲区**：围绕点的半径为缓冲距的圆形缓冲区；
   2) **线要素的缓冲区**：围绕线要素两侧距离不超过缓冲距的一系列长条形缓冲带；
   3) **面要素缓冲区**：围绕多边形边界内侧或外侧距离不超过缓冲距的面状区域；
   4) **复杂要素的缓冲区**：由复杂要素的单个目标缓冲区的并集组成的区域。
6. **点要素缓冲区的建立方法**：①直接绘制圆法；②圆弧步进拟合法。
7. **线要素缓冲区的建立方法**：以线状目标为参考轴线，以轴线为中心向两侧沿法线方向平移一定距离，并在线端点处以光滑曲线连接，所得到的点组成的封闭区域。
   1) **角平分线法**：在转折点处根据角平分线确定缓冲线的形状。



角平分线法的缺点：难以保证双线的等宽性、轴线转角尖锐的转折点将随缓冲距的增大迅速远离轴线。



2) **凸角圆弧法**：①在轴线的两端用半径为缓冲距的圆弧拟合；②在轴线转折点，判断该点的凹凸性，在凸侧用半径为缓冲距的圆弧拟合、在凹侧用与该点关联的两缓冲线的交点为对应缓冲点。



凸角圆弧法的优点：凸侧的缓冲线与轴线等宽，而凹侧的对应缓冲点位于凹角的角平分线上，最大限度地保证缓冲区边界与轴线的等宽关系。

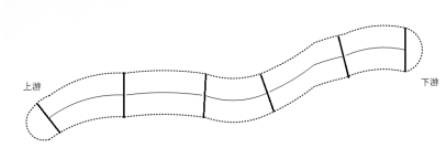
面要素缓冲区的建立方法：基本思路与线要素缓冲区生成算法基本相同。缓冲区分内侧缓冲区和外侧缓冲区，同一面状目标的内外侧缓冲区宽度可以不一样。

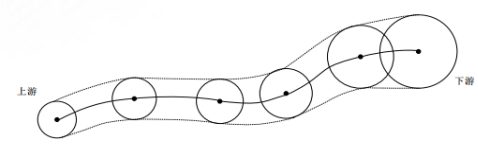
9. **动态缓冲区 VS 静态缓冲区**：

静态缓冲区：指空间目标对邻近对象的影响呈现单一的距离关系；

动态缓冲区：指空间目标对邻近对象的影响呈现不同强度的扩散或衰减关系。

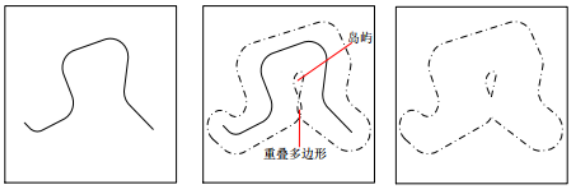
动态缓冲区生成算法：①基于线目标，用分段处理的办法分别生成各分段的缓冲区，然后将各分段缓冲区光滑连接；②基于点目标，用逐点处理的办法分别生成沿线各点的缓冲圆，然后求出缓冲圆序列的两外切线，所有外切线相连。





10. **缓冲线自相交问题：**当轴线的弯曲空间不能容许缓冲区边界自身无压覆地通过时，缓冲线将产生自相交现象，并形成多个自相交多边形：

    岛屿多边形——保留；重叠多边形——删除



11. **叠置分析**：在统一的空间坐标系下，将同一地区的两个或两个以上的地理要素图层进行叠置，产生空间区域的多种属性特征的分析方法。

    叠置分析的类型（6 种）：①点与点的叠置关系；②点与线的叠置关系；③点与面的叠置关系；④线与线的叠置关系；⑤线与面的叠置关系；⑥面与面的叠置关系。

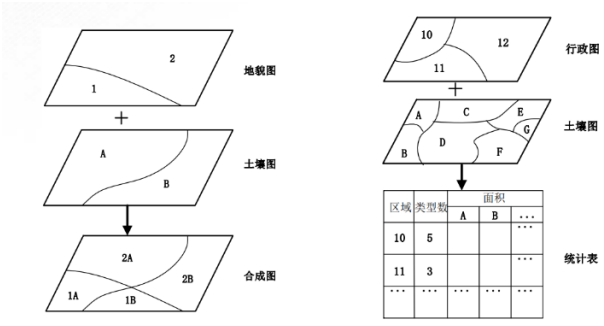
    1) **点与点、线、面的叠置、线与线的叠置：**可以用于建立新的属性：原属性 + 落在那个多边形的目标标识 + 从多边形属性表中提取一些附加属性。

    2) **多边形与多边形的叠置**：同一地区、同一比例尺的两组或两组以上的多边形要素进行叠置。

        a) **相关定义：**①多层叠置：两两叠置后再与第三层叠置。②本底多边形：被叠置的多边形为本底多边形。③上覆多边形：用来叠置的多边形为上覆多边形。④新多边形：叠置后产生具有多重属性的新多边形。

        b) **实现方法（3 步走）：**①将两层多边形的边界全部进行边界求交运算和切割；②根据切割的弧段重建拓扑关系；③判断新叠置的多边形分别落在原始多边形层的哪个多边形内，建立叠置多边形与原多边形的关系，如果必要再抽取属性。

        c) **两种基本处理方法：**①地图内容的合成叠置：根据多边形边界的交点建立具有多重属性的多边形；②地图内容的统计叠置：进行多边形范围内的属性特性的统计分析。



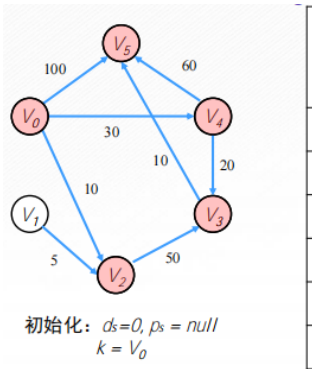
➢ 网络分析

1. **网络分析**：对地理网络和城市基础设施网络等网状事物及它们的相互关系和内在联系进行地理分析和模型化。
2. **网络分析中的常见问题**：
   1) **地址匹配（地址编码，Geocoding）**：基于空间定位技术的一种编码方法，提供一种把描述成地址的地理位置信息转换成地理坐标的方式。
   2) **路径分析**：最短路径、N 条最佳路径分析、静态求最佳路径、动态最佳路径分析
   3) **资源分配问题/资源分配模型**：由中心点（分配中心）及其状态属性和网络组成。
   4) **最佳选址**：指在一定约束条件下、在某一指定区域内选择设施的最佳位置，本质上是资源分配分析的延伸。
3. **最短路径问题**：d_k是从v_1到v_k的最短路径，设该路径的最后一段弧为(v_j, v_k)，w_{jk}为v_j到v_k的权值，由局部与整体的关系，该路径的前一段，即v_1到v_j的路径也必须为从v_1到v_k的最短路径。**最佳路径方程如下所示**：

$$\begin{cases} d_1 = 0 \\ d_k = \min(d_j + w_{jk}), j, k = 2,3,\dots,p; k \neq j \end{cases}$$

4. **迪杰斯特拉 Dijkstra 算法（最短路径算法）**：按路径长度递增的次序产生最短路径的算法。解决有向图中单个源点到其它顶点的最短路径问题。

   算法思想：
   1) 设$G = (V, E)$是一个带权有向图，把图中顶点集合 V 分成两组，第一组为已求出最短路径的顶点集合（用 S 表示，初始时 S 中只有一个源点，以后每求得一条最短路径 ，就将加入到集合 S 中，直到全部顶点都加入到 S 中，算法就结束了），第二组为其余未确定最短路径的顶点集合（用 U 表示），按最短路径长度的递增次序依次把第二组的顶点加入 S 中。
   2) 在加入的过程中，总保持从源点 v 到 S 中各顶点的最短路径长度不大于从源点 v 到 U 中任何顶点的最短路径长度。此外，每个顶点对应一个距离，S 中的顶点的距离就是从 v 到此顶点的最短路径长度，U 中的顶点的距离，是从 v 到此顶点只包括 S 中的顶点为中间顶点的当前最短路径长度。



终点	从V_0到各终点的距离值和最短路径的求解过程			
	i=1	i=2	i=3	i=4
V_1	∞	∞	∞	∞
V_2	(10, V_0)			
V_3	∞	(60, V_2)	(50, V_4)	
V_4	(30, V_0)	(30, V_0)		
V_5	(100, V_0)	(100, V_0)	(90, V_4)	(60, V_3)
V_j	V_2	V_4	V_3	V_5
S	(V_0, V_2)	(V_0, V_2, V_4)	(V_0, V_2, V_3, V_4)	(V_0, V_2, V_3, V_4, V_5)

① 对每个点进行设置，并对 V0 点进行标记

检测从V_0点到与之直接连接的未标记点之间的距离：

$$d(V_0,V_2)=10\,,d(V_0,V_4)=30\,,d(V_0,V_5)=100$$

得到：$S=\{V_0,V_2\}$，对V_2进行标记，$V_j=V_2$

③ 检测经过标记点V_0和V_2到与之直接连接的未标记点之间的距离：

$$d(V_0,V_2,V_1)=\infty\,,d(V_0,V_2,V_3)=60\,,d(V_0,V_4)=30\,,d(V_0,V_5)=100$$

得到：$S=\{V_0,V_2,V_4\}$，对V_4进行标记，$V_j=V_4$

④ 检测经过标记点V_0,V_2和V_4到与之直接连接的未标记点之间的距离：

$$d(V_0,V_2,V_3)=60\,,d(V_0,V_4,V_3)=50\,,d(V_0,V_5)=100\,,d(V_0,V_4,V_5)=90$$

得到：$S=\{V_0,V_2,V_4,V_3\}$，对V_3进行标记，$V_j=V_3$

⑤ 检测经过标记点V_0,V_2,V_3和V_4到与之直接连接的未标记点之间的距离：

$$d(V_0,V_5)=100\,,d(V_0,V_4,V_5)=90\,,d(V_0,V_4,V_3,V_5)=60$$

得到：$S=\{V_0,V_2,V_4,V_3,V_5\}$，对$V_5$进行标记，$V_j=V_5$

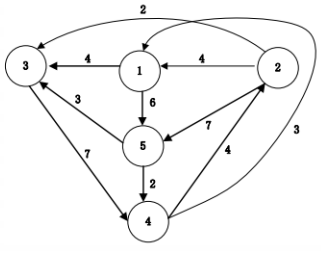
⑥ 接着对最后一个节点V_1进行标记，得到：$S=\{V_0,V_2,V_4,V_3,V_5,V_1\}$

5. **弗洛伊德 Floyd 算法（多点对最短路径算法）**：求解网络系统中多点对乃至所有结点对之间的最短路径。【也可以重复多次执行迪杰斯特拉算法】

$M(k)[i,j]$：由节点 i 到节点 j 经过节点序号不大于 k 的最短路径

$$M(k)[i,j]=\min\{M(k-1)[i,j],M(k-1)[i,k]+M(k-1)[k,j]\}$$

$M(1)[1,2]=M(2)[1,2]=M(3)[1,2]=\infty$;
$M(4)[1,2]=\min\{M(3)[1,2], M(3)[1,4]+M(3)[4,2]\}=\min\{\infty, 15\}=15$;
$M(3)[1,4]+M(3)[4,2]=\{1\text{-}3\text{-}4\}+\{4\text{-}2\}=11+4=15$
1 - 3 – 4 - 2
$M(5)[1,2]=\min\{M(4)[1,2], M(4)[1,3]+M(4)[3,2], M(4)[1,4]+M(4)[4,2], M(4)[1,5]+M(4)[5,2]\}$
$=\min\{15,15,15,12\}=12$
$M(4)[1,5]+M(4)[5,2]=\{1\text{-}5\}+\{5\text{-}4\text{-}2\}=6+6=12$



1到2的最短路径长度为12，路径为1-5-4-2

6. **次最短路径求解算法**：两点之间的次最短路径、第 3 短路径，…，第 k 短路径。

   基本思路：

   求出第 1 最短路径 P1 之后，用枚举法求出与 P1 有尽可能多公共边的次最短路径 P2。

   具体步骤：

   1) 假定第 1 最短路径 P1 包含了 n 条有向弧，每次删除其中的一条弧，即得到 n 个与原来只有一弧之差的新的网络。
   2) 按原最短路径算法分别求解这 n 个新网络的最短路径，然后比较这 n 条最短路径，其中最短的那条即为所求的次最短路径。
   3) 依此进行，可以分别求出第 3 短路径，…，第 k 短路径。

7. **最优路径分析**：指网络两结点之间阻抗最小的路径

   1) 基于单因素考虑的最优路径：时间最短、费用最低、风景最好、路况最佳、过桥最少、收费站最少、经过乡村最多等。
   2) 基于多因素综合考虑的最优路径：风景最好且经过乡村较多，或时间短、路况较佳、且收费站最少等。

8. **最大容量路径分析：**设网络D(V,E,W)中任意一条路径 P 的容量定义为该路径中所有弧的容量C_{ij}的最小值，即$C(P)=\min_{e_{ij}\in E(P)}(C_{ij})$，则网络D(V,A)中所有$(V_i,V_j)$路径中的容量最大的路径即为$(V_i,V_j)$的最大容量路径。

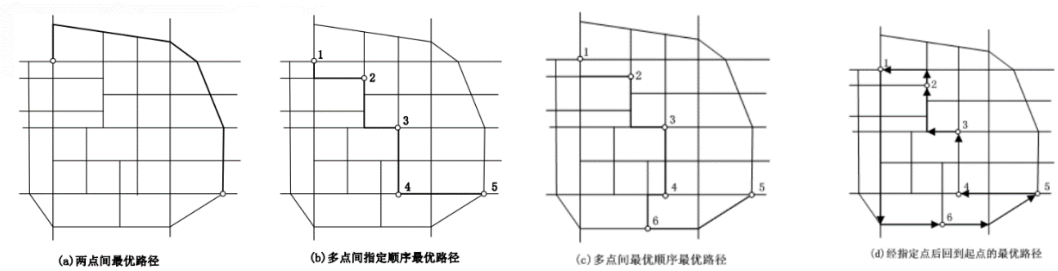
9. **时间最优路径分析**：

   将网络中的每条边或弧的权值定义为通过该边或弧的时间，就可以求出时间最优路径

   道路$u=\{e_1,e_2,\dots,e_k\}$的时间长度：u 上所有边的时间长度之和

   - **举一反三：道路$u=\{e_1,e_2,\dots,e_k\}$的费用**：u 上所有边的费用之和

10. **最优路径的求解的多种形式**：①两点间最优路径；②多点间指定顺序的最优路径；③多点间最优顺序最优路径；④经指定点回到起点的最优路径。



➢ 空间社会网络分析

1. **社会网络(Social Network，也称社交网络)**：是指社会个体成员之间因为互动而形成的相对稳定的关系体系，是由许多节点构成的一种社会结构，节点通常是指个人或组织。
2. **空间交互网络**：人、商品和信息等在地点之间流动而形成的嵌入在空间中的有向流网络。【典型的空间交互网络：国际贸易网络、人口迁移网络、人群出行网络、电话通信网络等】
3. **网络科学发展的三个时期**：
   1) **规则网络理论阶段**：得益于图论和拓扑学等应用数学的发展
      ✓ 四个经典图论问题：①哥尼斯堡七桥问题、②哈密顿问题、③四色猜想、④旅行商问题（又称货担郎问题、中国邮路问题）
   2) **随机网络理论阶段**：用相对简单的随机图描述网络。
      ✓ 随机网络：指由 N 个节点构成的图中以概率 P 随机连接任意两个节点而形成的网络
   3) **复杂网络理论阶段**：具有不同于规则网络和随机网络的特性的网络。
      ✓ 两大理论：①小世界理论（六度分割理论）：你和任何一个陌生人之间的间隔不会超过 6 个人。②无标度特性：如幂律分布、二八定律等。
4. **复杂网络模型**：
   1) **小世界网络（Small-world Network）**：即使网络规模很大，网络中任意两个节点间能通过较短的路径达成连接。
   2) **无标度网络（Scale-free Network）**：较少的节点掌握大量的资源（强者愈强、弱者愈弱）。
   3) **层次网络（Hierarchical Network）**：相对独立的模块组合相连，内部连接紧密，外部连接疏松。
5. **复杂网络的统计特性**：
   1) **度**：与节点i相连的节点个数。
   2) **度分布**：度的概率分布函数。
   3) **聚类系数**：要素集聚在一起的紧密程度（节点间的紧密程度）。
   4) **平均路径长度**：任意两点间最短路径的平均值（网络的平均通行效率）。

5) **介数**：经由该节点（边）的最短路径的个数。

6) **紧密度**：与其他节点的接近程度（当前节点至其它节点最短距离的倒数和）

第五章·人群活动分析及轨迹挖掘

➢ 城市人群活动

1. **城市人群活动的分类**：

   1) **现实空间活动**：工作、通勤、购物、休闲娱乐及其派生出来的交通出行活动等。

      - 受时空约束明显，呈现较强的时间-空间分异特性。

   2) **网络空间活动**：网络通信、网络购物、远程工作、在线休闲娱乐、在线自主学习

      - 突破距离，很强的自由性、开放性和自主性，活动多样与虚拟化等。

   3) **社交空间活动**：QQ、微信、微博、朋友圈、推特、分享体验等

      - 与现实与网络空间密不可分，呈现时空联动效应，表现群体化、多元化、多层次特征。

➢ 时间地理与时空 GIS

1. **时间地理**：研究物质环境中限制人们行为的制约条件，来说明人的空间行为。
2. **生命路径**：表示个体生命时间内位置信息的连续变化情况，为表示和研究相应的人文现象提供了一个十分直观、有效的方法。
3. **Space-time Path**：每个人在某时某地只能从事一种行为，时间和空间也是有限的资源
4. **可达性**：一种用来表征某个地点被接近的容易程度的属性，同时也被认为是用来表征人们到达一些潜在目的地进行某些活动的容易程度。

   1) **基于地点的（place-based）可达性**：表征某个地点被接近的容易程度的属性；

   2) **基于个人的（personal-based）可达性**：表征人们到达一些潜在目的地进行某些活动的容易程度。

➢ 人群活动群体特征分析

1. **人群活动群体特征**：移动性(Mobility)、聚散特性、功能性、适应性。
2. **移动性（Human mobility）**：表示人类个体/群体在地理空间中具有特定意义的“移动”所隐含的社会系统要素时空分布与演化规律。
3. **移动性的几种模型**：①重力模型(Gravity model)、②辐射模型(Radiation model)、③连

续随机游走模型(Continuous-time random walk)、④移动性 Markov 链模型(Mobility Markov Chain)、⑤T-模式树模型(Trajectory-pattern tree building)、⑥（卷积）神经网络模型(Recurrent) (Neural Network)、⑦注意力卷积网络模型(Attentional Recurrent Networks)

4. **城市功能区：**土地使用功能、使用强度、土地利用方向、基准地价大体一致的区域，其集约利用程度和使用潜力也基本相同。城市同种土地利用在空间上集中而形成集聚效应。
5. **人群移动模式与城市空间可达性的适应性分析**：计算相同出行网络的人群的适应性均值，表示此类人群的移动模式与城市空间适应度。
6. **人群交互模式与城市多中心结构适应性分析**：①规划中心体系结构与人群交互适应性指标、②城市发展轴带与城市人群交互适应性度量指标、③城市组团结构与人群交互适应性性度量指标。

➢ 轨迹分析与挖掘

1. **轨迹（Trajectory）：**移动对象的位置和时间的记录序列。

   ☆时空轨迹是时间到空间的映射：给定某一时刻 t，通过该函数 O 得到 t 时刻该对象所处的 d 维空间中的位置。

2. **轨迹数据预处理**：
   1) **轨迹匹配**：将轨迹点与浮动车行驶的路段进行匹配。
   2) **上下车点提取**：将车辆原始轨迹数据按照车辆 ID 和时间排序，“重车”和“空车”状态改变的轨迹点即为上下车点。
   3) **数据格式转换（txt 转 shp）**：轨迹数据的原始数据格式为 txt，一天内所有的轨迹信息为一份 txt 文件，每份 txt 内的轨迹信息按行存储，每一列是一个轨迹字段，每一行即为一个轨迹点。
   4) **异常数据清理**。
3. **热点与热点分析**
   1) **热点：**集中了较高热度的城市活动区域，吸引居民频繁出行的区域，是城市活力和功能运转的表征。
   2) **轨迹聚类：**利用轨迹聚类方法（K 均值聚类、时空数据场聚类等），从乘客上下车点中探测城市热点区域并分析其时空分布模式。

3) **城市热点区域时空分析：**可以从不同日期、时段等多方面进行热点区域分析。

4) **热点交互网络分析**：以热点区域为节点，乘客在热点区域之间的上下车关系为边，构建一张有向加权的热点区域交互网络。

5) **社区或社团：**是指网络中的节点内聚子图，子图内部的节点间存在较多的连接，不同子图的节点间连接相对稀少。选用随机游走模型对三天的热点网络进行社团探测，来分析热点区域之间交互的抱团性。

4. 异常轨迹检测与分析

5. **拥堵网络交互分析：**在交通网络中，车流可以分为自由流（Free Flow）和拥堵流(Congested Flow)两种类型，建立城市拥堵预警模型，为城市管理、交通规划等部门提供决策与支持，能在一定程度上缓解城市的交通拥堵问题。

1) **自由流：**在不受限制的条件下，车辆可以以任意速度进行行驶，分布特征是离散

2) **拥堵流：**车辆的行驶速度相对较低，甚至为零，分布特征是核聚，产生原因是网络瓶颈。

3) **拥堵复杂网络易崩溃结构分析：**易崩溃结构分布分析反应了拥堵网路的不稳定，是指比较容易发生拥堵的交叉口

6. **历史人物轨迹及社交关系分析**:

历史人物的轨迹数据：记录了历史人物在不同时间点的位置，隐含了空间地域上的活动、个体与社会的交互关系等社会属性。

第六章 · 三维分析与三维建模

➢ 三维地形模型与特征量算

1. **数字地面模型（DTM，Digital Terrain Model）：**描述地球表面形态多种信息空间分布的有序数值阵列。

➔数学角度：用二维函数系列取值的有序集合表示

$$K_p = f_k(U_p, V_p), k = 1,2,\dots,m;\ p = 1,2,\dots,n$$

其中K_p表示第 p 号地面点上的第 k 类地面特性信息的值；(U_p, V_p)表示第 p 号地面点的二维坐标；m 表示地面特性信息类型的数目；n 表示地面点的个数。

2. **数字高程模型（DEM，Digital Elevation Model）：**描述地球表面地面高程信息的有序数值阵列

➔DEM 是 DTM 的一个特例或者子集，数学角度：当 m=1 且f_1为地面高程的映射时

$$K_p = f_1(U_p, V_p), k = 1;\ p = 1,2,\dots,n$$

3. **数字表面模型（DSM，Digital Surface Model）：**包含了地表建筑物、桥梁和树木等高度的三维地形模型。

4. **DEM 的表示方法**：

1) **数学方法**：

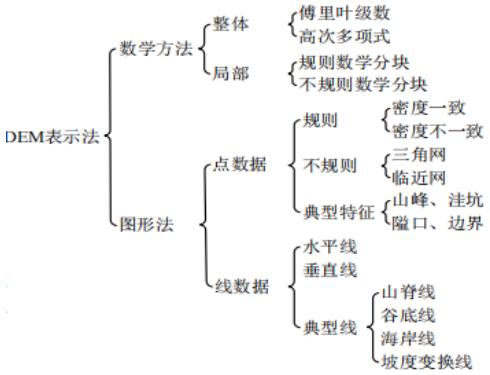
①**整体拟合**：将区域中所有高程点的数据用傅里叶级数、高次多项式、随机布朗运动函数等统一拟合高程曲面。

②**局部拟合**：把地面分成若干块，每一块用一种数学函数拟合。

2) **图形法**：

①**线模式：**用离散的地形特征模型表示地形起伏，例如等高线、山脊线、谷底线、海岸线和坡度变换线。

②**点模式：**用离散采样数据点建立 DEM（规则格网模式 Grid、不规则模式 TIN）



5. **规则格网模型 Grid 与不规则三角网 TIN：**

1) **规则格网：**将区域空间切分为规则的格网单元，每个格网单元对应一个数值，且每一个格网点与相邻格网点之间的拓扑关系都可以从行列号中反映出来。

2) **不规则三角网：**将采集的地形特征点根据一定的规则构成覆盖整个区域且不重叠的一系列三角网。

3) **Grid 与 TIN 各自的优缺点**：

Grid：①结构简单，易于计算机处理；②可以很容易地计算等高线、坡度坡向、山坡阴影和自动提取流域地形；③对于地形简单的地区存在大量冗余数据；④如果不改变格网大小无法使用于地形起伏差别较大的地区，且栅格过于粗略不能精确表示地形的关键特征。

TIN：①从等高线数据选取重要点构成 TIN 并生成的规则格网，在数据量相同的情况下，TIN 数据具有最小的中误差；②TIN 与数字正射影像（DOM）的叠加比规则格网更好；③当采样点数量减少时，规则格网模型的质量比 TIN 模型降低的速度更快，但随着采样点或数据密度的增加，二者差别会逐步减小。

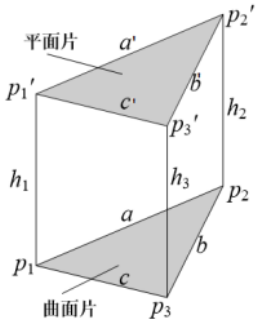
6. 三维空间特征量算——表面积计算

1) **三角形格网上的表面积计算**:

将三角形曲面片$P_1P_2P_3$使用一次多项式拟合得到平面片$P_1'P_2'P_3'$，计算平面片面积

$$a'=\sqrt{a^2+(h_1-h_2)^2}\ ;\ b'=\sqrt{b^2+(h_2-h_3)^2}\ ;\ c'=\sqrt{c^2+(h_3-h_1)^2}$$

$$\rightarrow\ S=\sqrt{P(P-a')(P-b')(P-c')}\ ;\ P=\frac{a'+b'+c'}{2}$$



2) **正方形格网上的表面积计算**：①曲面拟合重积分方法；②分解为三角形的方法

a) **曲面拟合重积分方法**：先用二次抛物面逼近面积计算函数（近似计算，又称辛卜生方法），进而将抛物面的表面积计算转换为函数值计算。

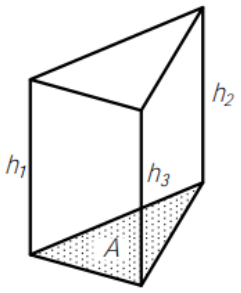
➔空间单值曲面$Z=f(x,y)$的面积：$S=\iint\sqrt{1+f_x^2+f_y^2}\,dxdy$

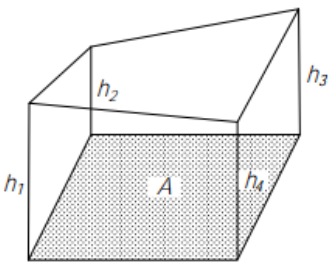
b) **分解为三角形的方法**：将正方形格网 DEM 的每个格网分解为三角形，分别计算分解的三角形的面积，然后累加即得到正方形格网 DEM 的面积。

7. 三维空间特征量算——体积计算

1) 基于三角形格网的体积计算：$V=S_A\times\frac{h_1+h_2+h_3}{3}$

2) 基于正方形格网的体积计算：$V=S_A\times\frac{h_1+h_2+h_3+h_4}{4}$



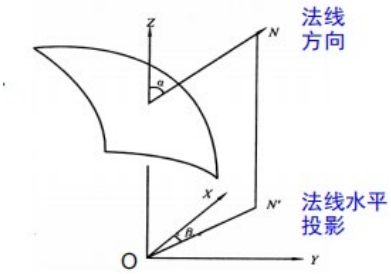


➢ 地形分析

1. **坡度与坡向：**

   1) **坡度：**某点在曲面上的法线方向与垂直方向的夹角，是地面特定点高度变化比率的度量 ➔ 反映斜坡的倾斜程度

   2) **坡向：**法线的正方向在平面上的投影与正北方向的夹角➔反映斜坡所面对的方向



2. **坡度和坡向计算方法：**①基于规则格网；②基于不规则三角网

3. **基于规则格网 GRID 的坡度坡向计算**

原理：由单元标准矢量的倾斜方向和倾斜量，计算每个单元的坡度和坡向。

Tips：标准矢量：垂直于格网单元的有向直线

➔实际计算时，通常用 3×3 的移动窗口来计算中心单元的坡度和坡向

① **Ritter 算法：**利用与中心点单元 e 直接相邻的 4 个单元中心点的高程值计算

➔**坡度：** $S_e = \frac{\sqrt{(e1-e3)^2+(e4-e2)^2}}{2d}$ **坡向：** $D_e = \arctan\left(\frac{e4-e2}{e1-e3}\right) + 90°$

➔d 表示单元大小，$(e1-e3)$表示 x 方向的高差，$(e4-e2)$表示 y 方向的高差

	e_2	
e_1	e	e_3
	e_4	

② **Horn 算法：**利用与中心单元相邻的 8 个相邻单元来计算，直接邻接单元 $(e2, e4, e5, e7)$的权值为 2，其它 4 个单元$(e1, e3, e6, e8)$的权值为 1。【ArcGIS 采用】

➔坡度：$S_e = \frac{\sqrt{((e1+2e4+e6)-(e3+2e5+e8))^2+((e6+2e7+e8)-(e1+2e2+e3))^2}}{8d}$

➔坡向：$D_e = \arctan\left(\frac{(e6+2e7+e8)-(e1+2e2+e3)}{(e1+2e4+e6)-(e3+2e5+e8)}\right)$

e_1	e_2	e_3
e_4	e	e_5
e_6	e_7	e_8

4. 基于不规则三角网 TIN 的坡度坡向计算

$$\overrightarrow{E_1E_2} = (x_2 - x_1, y_2 - y_1, z_2 - z_1) \quad ; \quad \overrightarrow{E_1E_3} = (x_3 - x_1, y_3 - y_1, z_3 - z_1)$$

➔ 标准向量为$\overrightarrow{E_1E_2}$与$\overrightarrow{E_1E_3}$的向量积，即：

$$\vec{n} = \overrightarrow{E_1E_2} \times \overrightarrow{E_1E_3} = \begin{bmatrix} i & j & k \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{bmatrix}$$

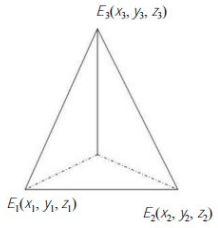
$\vec{n}$的三个分量分别为：

$$n_x = (y_2 - y_1)(z_3 - z_1) - (y_3 - y_1)(z_2 - z_1)$$

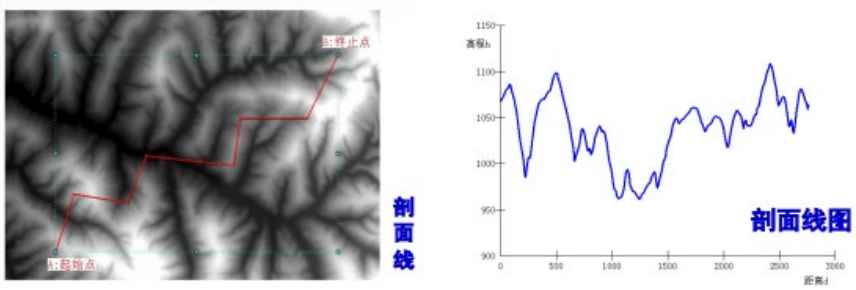
$$n_y = (z_2 - z_1)(x_3 - x_1) - (z_3 - z_1)(x_2 - x_1)$$

$$n_z = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)$$

➔ **坡度：** $S = \frac{\sqrt{{n_x}^2 + {n_y}^2}}{n_z}$ **坡向：** $D_e = \arctan\left(\frac{n_x}{n_y}\right)$



5. **剖面分析：** 以数字高程模型(DEM)为基础构造某一方向的剖面，以线代面，概括研究区域的地势、地质和水文特征。



6. **地形剖面线：** 根据所选剖面与数字地形图上地形表面的交点反应地形的起伏情况。

7. **地形剖面线的生成方法**：①基于规则格网的方法、②基于不规则三角网的方法。

8. **基于规则格网的剖面线生成算法**：

   1) 确定剖面线的起止点：由坐标确定，或用鼠标在三维场景中选择决定；

   2) 计算剖面线与经过网格的所有交点，内插出各交点的坐标和高程，将交点按离起始点的距离排序；

   3) 顺序连接相邻交点，得到剖面线；

   4) 选择一定的垂直比例尺和水平比例尺，以离起始点的距离为横坐标，以各点的高

程值为纵坐标绘制剖面图。

9. **基于不规则三角网的剖面线生成方法**：

1) 用剖面所在的直线与 TIN 中的三角面的交点得到剖面线；

2) 利用 TIN 中各三角形构建的拓扑关系快速找到与剖面线相交的三角面；

3) 进行交点的计算；

4) 以距离起始点的距离为横坐标，以各点的高程值为纵坐标绘制剖面图

10. **谷脊分析：**对地形的谷点和脊点进行分析。

➔**谷：**地势中相对最低点的集合；**脊：**地势中相对最高点的集合

11. 基于栅格 DEM 的谷脊点概略计算方法

设h_x为某点的高程值，则：

1) **当$(h_{i,j-1}-h_{i,j})\times(h_{i,j+1}-h_{i,j})>0$时**：

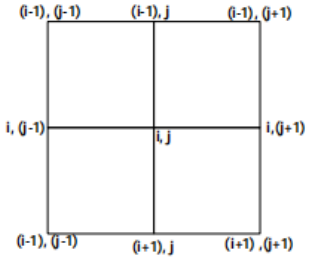
①若$h_{i,j+1}>h_{i,j}$，$h_{i,j-1}>h_{i,j}$ ➔$V_{R(i,j)}=-1$

②若$h_{i,j+1}<h_{i,j}$，$h_{i,j-1}<h_{i,j}$ ➔$V_{R(i,j)}=1$

2) **当$(h_{i-1,j}-h_{i,j})\times(h_{i+1,j}-h_{i,j})>0$时**：

①若$h_{i+1,j}>h_{i,j}$，$h_{i-1,j}>h_{i,j}$ ➔$V_{R(i,j)}=-1$

②若$h_{i+1,j}<h_{i,j}$，$h_{i-1,j}<h_{i,j}$ ➔$V_{R(i,j)}=1$



➔ 根据$V_{R(i,j)}$的值确定(i,j)是谷点、脊点或是其他点

$V_{R(i,j)}=-1$➔(i,j)为谷点；$V_{R(i,j)}=1$➔(i,j)为脊点；$V_{R(i,j)}=0$➔(i,j)为其他点

12. **基于曲面方程的谷脊特征精确分析**：由曲面拟合方程建立地表单元的曲面方程，确定曲面上各个插值点的极小值和极大值。

1) **脊点：**当插值点在两个相互垂直的方向上为极大值时，确定出脊点。

2) **谷点：**当插值点在两个相互垂直的方向上为极小值时，确定出谷点。

13. **水文分析**：由 DEM 数据对地表水流进行分析。（水文分析主要包括：无洼地 DEM 生成、汇流累积矩阵计算、水流长度计算、河网提取、流域分割）

➔数据基础：无洼地的 DEM（对源 DEM 数据进行洼地填充）；关键步骤：流向分析

14. **水流方向：**指水流离开此格网时的方向，通过计算中心格网与邻域格网的最大距离权落差（中心栅格与邻域栅格的高程差除以两栅格间的距离）确定的。

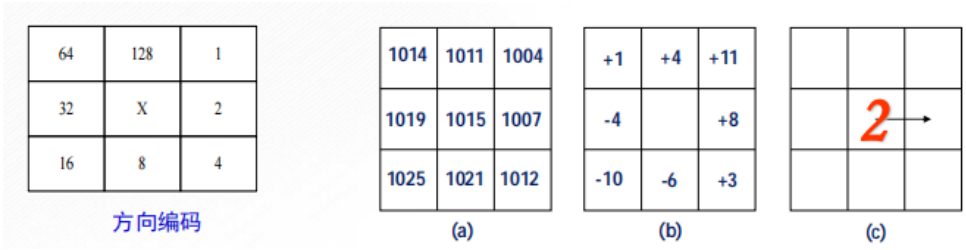
水流方向的方向编码：通过将格网 X 的 8 个邻域格网编码，水流方向便可以其中的一个值来确定。方向值以 2 的幂值指定，在后续处理中根据方向值相加结果可以确定相

加时中心格网的邻域格网状况。

两栅格间的距离：

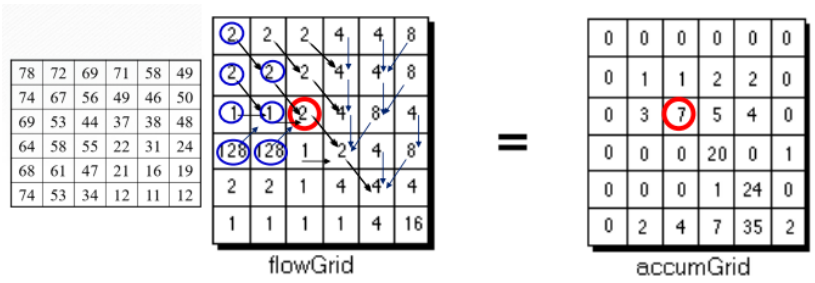
①方向数为 1,4,16,64，距离为栅格单元边长的$\sqrt{2}$倍；

②方向数为 2,8,32,128，距离为栅格单元的边长。



15. **汇流累积数值矩阵**：表示区域地形每点的流水累积量，基于水流方向数据计算得到。

计算汇流累积矩阵的基本思想：以规则格网表示的数字地面高程模型的每点都有一个单位水量，按照自然水流从高处往低处流的自然规律，根据区域地形的水流方向数据计算每点处所流过的水量数值，计算得到该区域的汇流累积量。



16. **水流长度**：地面一点沿水流方向到其流向终点的最大地面距离在水平面上的投影长度

17. **河网提取**：采用地表径流漫流模型

1) 在无洼地 DEM 上利用最大坡降法计算出每一个栅格的水流方向；

2) 计算每一个栅格在水流方向上累积的水量数值，即汇流累积量；

3) 假设每一个栅格携带一份水流，栅格的汇流累积量就代表着该栅格的水流量；

4) 当汇流量达到一定值的时候，就会产生地表水流，所有汇流量大于临界值的栅格就是潜在的水流路径，由这些水流路径构成的网络就是河网。

18. **水淹分析**：基于 DEM 数据对洪水淹没特性进行分析。

需要考虑的因素：洪水特性、受淹区域的地形地貌

19. **洪水淹没方式以及相关的问题**：

① **漫堤式淹没**：堤坝没有溃决，由于洪水水位过高导致的洪水从堤坝顶部进入淹没区

特定水位条件下：洪水会导致多大的淹没范围和多高的水深分布

② **决堤式淹没**：堤坝溃决，洪水从溃决处进入淹没区

➔某一洪量条件下：洪水可能造成多大的淹没范围和水深分布

20. 给定洪水水位的淹没分析的步骤

1) 确定 DEM 数据中的**洪水水源入口**

2) **淹没范围**：根据给定的洪水水位，从水源处开始进行格网连通性分析，所有能够与洪水水源入口处连通的格网单元构成洪水淹没的范围。

3) **淹没水深**：淹没范围内格网的水深（洪水水位减去格网单元高程值）。

21. **能够形成地表径流的地貌形态**：河流及洪水形成的山谷沟渠➔河流和山谷属于谷地地貌，可以通过山谷线来判断➔谷脊分析得到地表径流路径。

- 山谷线均由连续的局部极小值构成；
- 山谷线由其最高点(上游)往下游延伸的其它山谷线特征点的高程值应越来越小；
- 山谷线的终止条件：①连接另一条山谷线；②汇入湖泊海洋；③到达 DEM 边缘

22. 洼地连通分析

1) **河流沟谷本来就终止于该洼地**：①通过山谷线分析得到山谷线➔②根据水流方向直接往下游追踪➔③最后到与该沟谷(或河流)连接的洼地，得到两者的连通关系

2) **当被淹没的洼地水位到达一定程度时，水从洼地边缘漫出，流向其它较低地区**：通过分析找到洼地边缘和溢口，并判断流水的流向。

3) **洼地连通分析的关键在于寻找洼地边缘**：射线法和扩散法(子蔓延法)

$V_{R(i,j)} = 1$的点视为洼地边缘点

a) **射线法**：从平行线和铅垂线两个方向扫描洼地边缘点

b) **扩散法**：将洼地底点中的一个点作为种子点向周围相邻的 8 个方向扩散

4) **洼地溢口点**：该洼地边缘点中高程值最小的点。

23. **给定洪量的淹没分析的步骤**：

1) 计算给定水位条件下淹没区域的容积，将容积与洪量相比较

2) 利用二分法等逼近算法找出与洪量最接近的容积，容积对应的淹没范围和水深分布即最后的分析结果。

➢ 三维建模与可视分析

1. **三维空间数据模型**：是研究三维空间的几何对象的数据组织、操作方法以及规则约束条件等内容的集合。

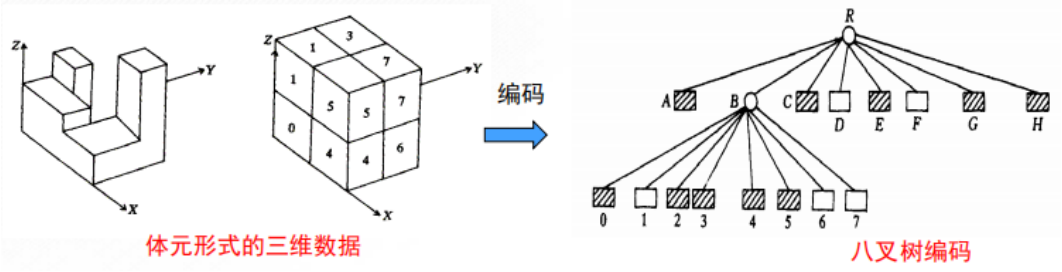
2. **常用的三维空间数据模型**

1) 面模型(facial model)：侧重于三维空间表面的表示，如：地形表面，地质层面等；

2) 体模型(volume model)：侧重于三维空间体的表示，如：水体、建筑物等；

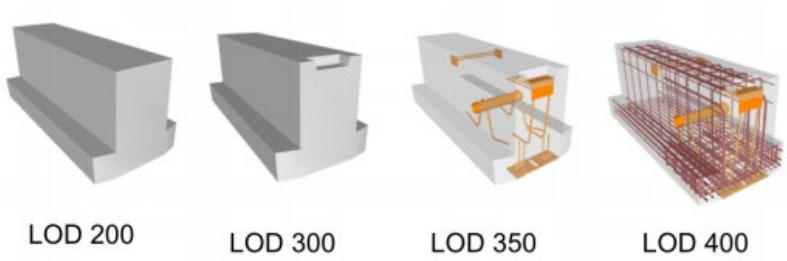
3) 混合模型(mixed model)：包含了面模型和体模型。

3. **三维场景的数据模型——八叉树模型**（八叉树数据结构是三维栅格数据的压缩形式，是二维栅格数据中的四叉树在三维空间的推广）



- **小圆圈**表示立方体未被某目标填满，或者说它包含有多个目标，需要继续划分；
- **有阴影线的小矩形**表示该立方体被某个目标填满；
- **空白的小矩形**表示该立方体中没有目标

4. **细节层次(Level of Detail，LOD)**：在减少所需要绘制的数据内容和几何细节时，保证观察者无法感知场景细节程度的降低，其主要的实现方法与相关因素，包括剪切、距离、大小、偏心率（物体偏离视觉中心的距离）、物体运动速度等。



5. **城市三维场景建模的主要思想**：BIM（建筑信息模型，LOD-4）、CAD 模型（LOD-3）、3DGIS 模型（LOD-2，LOD-1）三位一体。

6. **BIM(Building Information Modeling)的几何模型的形式**：①B-rep（边界表示）；②CSG （constructive solid geometry 结构实体几何）；③Sweep volumes（放样体）。

7. **CityGML 模型**：基于 XML 格式的用来表现城市三维对象的通用信息模型

- CityGML 文件可以同时包含每个对象的多个细节层次

➔ **LOD0（地域模型 Regional model)**：2.5D 数字地形图

（城市/场地模型 City/ Site model)：没有屋顶结构的“楼块模型”

➔ **LOD2（城市/场地模型 City/ Site model）**：包含贴图和楼顶结构的粗模

➔ **LOD3（城市/场地模型 City/ Site model）**：包含更多细节的建筑模型

➔ **LOD4（室内模型 Interior model）**：可以“步行进入”的建筑模型

8. 在 3DGIS 的三维自适应可视化表示特别是模型视图和世界视图两种透视表示中，显示效果的逼真性和交互响应的实时性是两个重要的指标。

9. **三维场景可视化技术：**①LOD；②背面剔除；③遮挡剔除。

10. **虚拟现实技术：**是指借助计算机系统及传感器技术生成三维环境，创造出一种崭新的人机交互方式，通过调动用户的各种感官(视觉、听觉、触觉、嗅觉 等)来享受更加真实的、身临其境的体验。

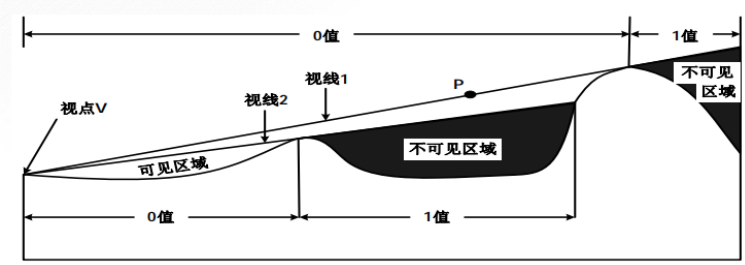
    - **基本特征：**虚拟性(Imagination)、沉浸感(Immersion)、实时交互性(Interactivity)

11. **三维场景可视分析与应用**：

    1) 屏幕坐标到三维空间坐标的转换；

    2) 空间射线与三角形交点计算（计算射线与三角形所构成平面的交点➔判断交点是否在三角形内）；

    3) 通视分析：可简化为 DEM 格网与某一地形剖面线(视线)的相交问题



    4) 日照分析与阴影图

    5) 三维叠置分析

12. **三维建模的工具：**用 CAD 或 3DMAX、Revit 软件等进行交互式构建。

13. **三维渲染引擎：**①OpenSceneGraph（用 C++）；②Web3D(用 Javascript)

第七章・探索性空间数据分析

➢ 一般统计分析

1. **GIS 属性数据的一般统计分析的过程**：先对数据进行描述性统计分析，以发现其内在规律➔再选择进一步分析的方法。

2. **描述性统计分析：**对调查总体所有变量的有关数据进行统计性描述，包括频数分析、数据的集中趋势分析、数据的离散程度分析、数据的分布、以及基本的统计图形。

3. **数据的频数分析**

   1) 频数：变量在各组出现或发生的次数称为频数

   2) 频率：各组频数与总频数之比叫做频率

   3) 频率分布图：表示各组的频率分布

   4) 频率直方图：若以纵轴表示频率，横轴表示分组，就可做出频率直方图，用以表示事件发生的概率和分布状况

4. **数据的集中趋势分析**

   1) **平均值**：是衡量数据的中心位置的重要指标，反映了一些数据必然性的特点。

      ➔算术平均值：将所有数据相加，再除以数据的总数目

      ➔加权算术平均值：每个数据乘其权值后相加，所得的和除以数据的总体权重数

      ➔调和平均值：各个数据的倒数的算术平均数的倒数，又称为倒数平均值

      ➔几何平均数：n 个数据连乘的积开 n 次方根

   2) **中位数**：一种反映数据的中心位置的指标，其确定方法是将所有数据以由小到大的顺序排列，位于中央的数据值就是中位数。

   3) **众数**：数据中发生频率最高的数据值。

   - 数据间差异程度较小➔平均值；数据之间差异程度大➔中位数或众数

5. **数据的离散程度分析：**反映数据之间的差异程度

   1) **方差/均方差：**离差平方和除以变量个数，即：$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

   2) **标准差：**方差的平方根，即：$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

   3) **极差：**一组数据中最大值与最小值之差

   4) **离差：**一组数据集中的各数据值与其平均数之差，即$d = x_i - \bar{x}$

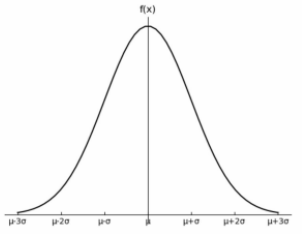
   5) **平均离差：**将离差取绝对值，然后求和，再取平均数，即$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

   6) **离差平方和**：对离差求平方和，即$d^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

6. **数据的分布分析**

   1) **正态分布：**用均值μ和方差σ^2两个数字特征表示，正态分布的定义如下：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$



- 用偏度和峰度两个指标来检查样本是否符合正态分布

  偏度：样本分布的偏斜方向和程度，是统计数据分布偏斜方向和程度的度量

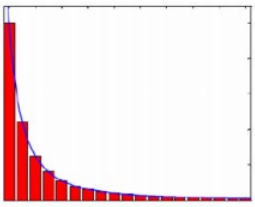
  峰度：样本分布曲线的尖峰程度，用来反映频数分布曲线顶端尖峭或扁平程度

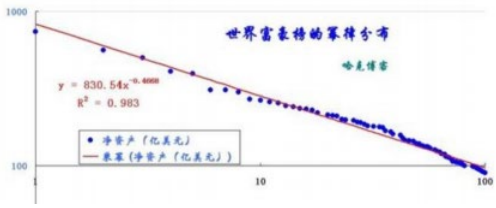
  如果样本的偏度接近于 0，而峰度接近于 3，可以判断总体的分布接近于正态分布

2) **幂律分布（长尾分布、重尾分布）**：分布函数$f(x)$与变量 x 之间是幂函数关系

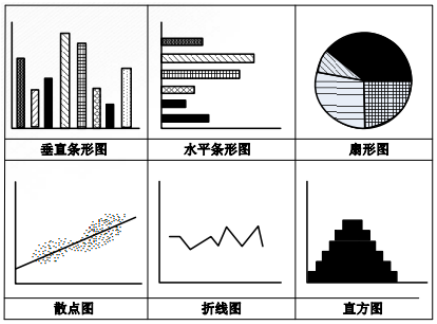
$$f(x) = kx^{-\gamma}\ (x > 0)$$

如果对幂律函数两边取对数，在双对数坐标系下，幂律分布呈直线，该直线的斜率为幂指数的负数：$\ln f(x) = \ln k - \gamma lnx$





7. **统计图表分析**：柱状图、扇形图、直方图、折线图和散点图等



探索性空间数据分析）

1. **空间统计：**空间统计是对具有空间分布特征数据的统计分析理论与方法。
2. **空间统计分析的主要类型**：①探索性空间数据分析、②空间点模式分析、③空间相关性分析、④地统计分析。
3. **探索性数据分析（EDA，exploratory data analysis）：**是不对数据总体做任何假设（或很少假设）的条件下识别数据特征和关系的分析技术。

   EDA 的特点：①对数据总体不作假设、②假设检验也经常被排除在外。

【核心在于“让数据说话”，在探索的基础上再对数据进行更为复杂的建模分析】

4. **探索性数据分析的基本方法**:

   1) **计算 EDA 方法：**从简单的统计计算到高级的用于探索分析多变量数据集中模式的多元统计分析方法。

   2) **图形 EDA 方法：**可视化的探索数据分析，包括：常用的图形方法有直方图(histogram)、茎叶图(stem leaf)、箱线图(box plot)、散点图(scatter plot)、平行坐标图(parallel coordinate plot)等。

5. **直方图 vs 茎叶图：**表述数据的分布信息

   1) **直方图：**二维统计图表，两个坐标分别是统计样本和该样本对应的某个属性度量

   2) **茎叶图：**又称“枝叶图”，将数组中的数按位数进行比较，将数的大小基本不变或变化不大的位作为主干(茎)，将变化大的位的数作为分枝(叶)，列在主干的后面，可以清楚地看到每个主干后面的数。

茎 | 叶	频数
0 | 1569	4
1 | 0569	4
2 | 24	2
3 | 1	1
4 | 016	3
5 | 257	3
6 | 0159	4
7 | 0159	4
8 | 59	2
9 | 124	3

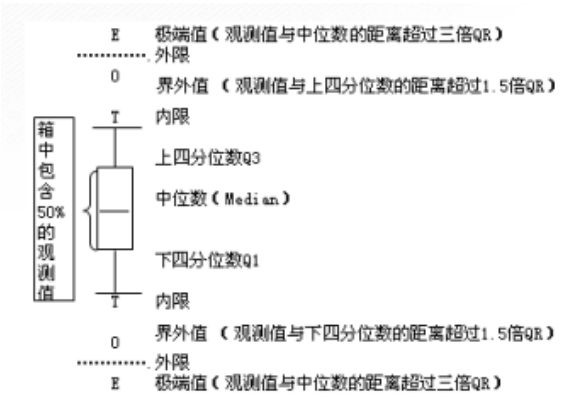
41, 52, 6, 19, 92, 10, 40, 55,
60, 75, 22, 15, 31, 61, 9, 70,
91, 65, 69, 16, 94, 85, 89, 79,
57, 46, 1, 24, 71, 5

- 茎叶图保留了原始资料的信息，直方图则失去了原始数据的信息
- 茎叶图中的数据可以随时记录、随时添加，但只便于表示两位有效数字的数据

6. **箱线图：**利用数据中的五个统计量（最小值、第一四分位数/下四分位数 Q1、中位数 F、第三四分位数/上四分位数 Q3、最大值）来描述数据，能够直观明了地识别数据集中的异常值。

   - 箱线图的绘制依靠实际数据，不需要事先假定数据服从特定的分布形式
   - 四分位距（QR）：上四分位数与下四分位数之间的间距，即上四分位数减去下四分位数。公式表示为：$QR = Q3 - Q1$

   画一个矩形盒，两端边的位置分别对应数据集的上下四分位数。在矩形盒内部的中位数位置画一条线段为中位线。在 Q3+1.5QR 和 Q1-1.5QR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称其为内限。在 Q3+3QR 和 Q1-3QR 处画两条线段，称其为外限。



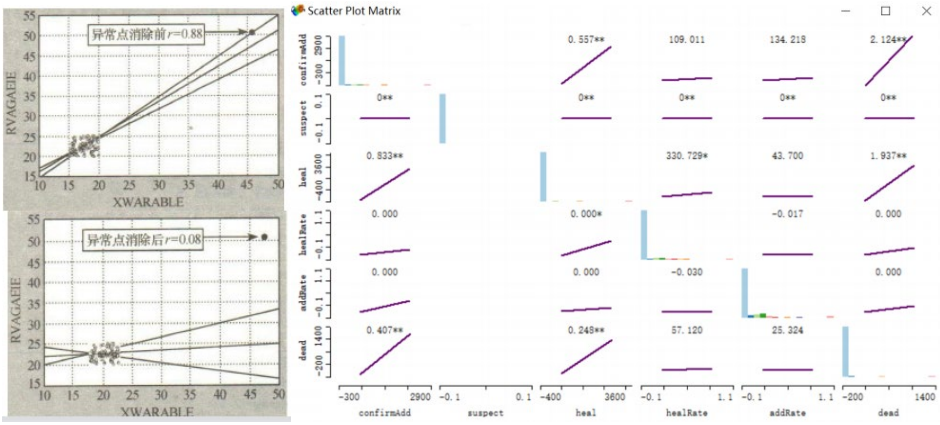
➔内限以外位置的点表示的数据都是**异常值**（x< Q1-1.5QR 或 x> Q3+1.5QR）

➔在内限与外限之间的异常值为**温和异常值**（Q1-3QR <x< Q1-1.5QR 或 Q3+1.5QR <x< Q3+3QR）

➔在外限以外的为极端**异常值**（x< Q1-3QR 或 x> Q3+3QR）

7. **散点图：**初步图示两个数据之间的关系，分析两个要素或变量之间关系

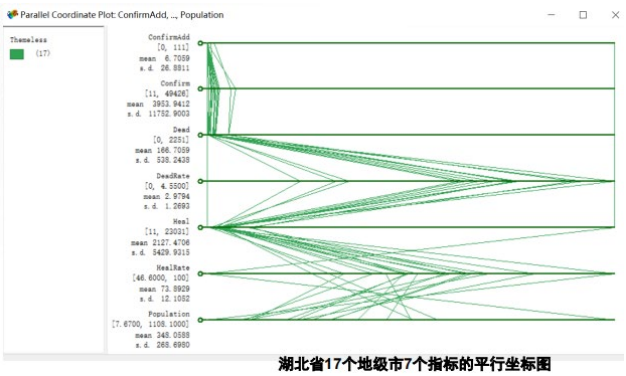
散点图矩阵：通过建立任意两个变量之间的关系的图形表示来获得相关信息和异常信息，相当于在由 m 个变量构成的矩阵中，用相应的两个变量之间的散点图替代矩阵中的元素构成的图形。



异常数据或者有着特别的价值，或者会引起错误的结果或判断 ➔ 剔除数据可能会对变量之间的关系产生很大的影响 ➔ 在回归模型建立之前通过散点图技术进行数据的探索性分析，有利于消除异常数据，寻找更为合理的关系或模式。

8. **平行坐标图：**在二维平面上表示高维空间中变量之间关系，为可视化地探索分析高维数据空间中的关系建立了可行的途径。

- 传统坐标系中所有的变量轴都是交叉的，而平行坐标系中所有变量轴都是平行的



湖北省17个地级市7个指标的平行坐标图

9. **探索性空间数据分析（ESDA）**：将数据的统计分析和地图定位紧密结合在一起，通过地理空间（地图表示）和属性空间（数据空间）的关联来凸显空间关系。
10. **ESDA 的主要作用：**①概括空间数据的性质；②探索空间数据中的模式；③产生和地理数据相关的假设；④在地图上识别异常数据的分布位置；⑤发现是否存在热点区域(hotspots)
11. **GIS 环境中 ESDA 的主要方法**：
    1) **动态联系窗口(dynamic linking windows)：**通过刷新技术将地理空间和属性空间的各种视图组合在一起，是一种交互式探索空间数据的选择、聚集、趋势、分类、异常识别的工具。

       动态联系窗口的特点：①在一种信息窗口中点击或选择，其它的信息窗口产生相应的响应，并高亮显示选中的信息。②ESDA 将多种可视化的数据分析工具和地图分析结合在一起，并提供了丰富的交互工具，不仅可以进行选择操作，而且能够进行改变数据参数等模式的探索。
    2) **刷新(brushing)技术**。
12. **ESDA 提供了两类统计分析方法**：①全局方法(global)：对所有实例的一个或多个属性数据进行处理、②局部方法(local)：对某个时段的数据子集进行统计分析。
13. **专题地图的数据分类方法：**等间隔、等范围、自然分割法、分位数分类、自定义等
    1) **等间隔分类**：假设分割之间的距离是相同的。
    2) **分位数分类**：将所有的观测数据按照相等的数量分配到每一个类中。

       自然分割分类方法：用户沿着数字线选择最大的分割，或在数据出现显著的空隙【基本思想：最小化数据集内部的变异、最大化类型间的差异（聚类）】

    - **分类的注意事项：**①包括所有范围的数据（最小和最大）；②使用不重叠的值和不空的类；③分类数量足够大，以避免牺牲数据的精确性；④划分数据集到合理的

等价的观测组中。

14. **地理空间(geographic space)：**由空间参考数据构成的坐标空间，使用地理坐标定义地理事物和现象，是地图形式的地理表示。

15. **数据空间(data space)：**地理实体属性所构成的空间，每一个点代表地理事物在数据空间中的位置。

第八章·地理相关分析方法

➢ 一般相关程度的度量方法

1. **相关：**描述两个或两个以上变数间相互关系是否密切。

2. **地理相关：**应用相关分析法研究各地理要素间的相互关系和联系强度的一种度量指标

3. **地理要素间的关系**

   1) **函数关系：**确定性的关系，这种关系在地理各要素间较少见，这是因为许多地理要素的变化具有随机性的缘故。

   2) **相关关系：**即要素间既存在密切的关系，但又不能由一个（或几个）要素（或变量）的值明确地求出另一个要素（变量）的值。

4. **相关方向**

   1) **正相关：**y 值随 x 的增加而变大或随 x 的减少而变小；

   2) **负相关：**y 值随 x 的增加而变小或随 x 的减少而增大。

5. **相关系数r_{xy}的计算**：

   对于两个要素 x 与 y，它们样本值分别为x_i和y_i，则他们之间的相关系数被定义为：

$$r_{xy}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{\sum x_i y_i-\frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum {x_i}^2-\frac{(\sum x_i)^2}{n}\right)\left(\sum {y_i}^2-\frac{(\sum y_i)^2}{n}\right)}}$$

   r_{xy}为两个要素 x 与 y 的相关系数，表示两个要素之间相关程度的统计指标

   r_{xy}的取值范围区间为：$[-1,1]$

   ①$r_{xy}>0$表示正相关；②$r_{xy}<0$表示负相关；③r_{xy}的绝对值越接近于 1，表示两要素的关系越密切； 越接近于 0，表示两要素的关系越不密切。

6. **相关系数的检验：**相关系数是根据要素之间的样本值计算出来，它随着样本数的多少或取样方式的不同而不同，因此它只是要素之间的样本相关系数，只有通过检验，才

能知道它的可信度。➔检验是通过在给定的置信水平下，查相关系数检验的临界值表来实现的。

7. 检验相关系数$\rho=0$（即两要素不相关）的临界值（r_α）表

	0.10	0.05	0.02	0.01	0.001
1	0.98769	0.99692	0.999507	0.999877	0.999998
2	0.90000	0.95000	0.98000	0.99000	0.999000
3	0.8054	0.8783	0.93433	0.95873	0.991160
4	0.7293	0.8114	0.8822	0.91720	0.97406
5	0.6694	0.7545	0.8329	0.8745	0.95074
6	0.6215	0.7067	0.7887	0.8343	0.92493
7	0.5822	0.6664	0.7493	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
9	0.5214	0.6021	0.6851	0.7348	0.8471
10	0.4973	0.5760	0.6581	0.7079	0.8233
11	0.4762	0.5529	0.6339	0.6835	0.8010
12	0.4575	0.5324	0.6120	0.6614	0.7800

对公式 $p=\{|r|>r_\alpha\}=\alpha$ 的解释：当所计算的相关系数r的绝对值大于在α水平下的临界值r_α时，两要素不相关（$\rho=0$）的概率只有α。

左边的f值称为自由度，其数值为$f=n-2$，这里为n样本数；上方的α代表不同的置信水平；表内的数值r_α代表不同的置信水平下相关系数$\rho=0$的临界值。

➔一般而言，当 $|r|<r_\alpha$ 时，则这时的样本相关系数就不能反映两要素之间的关系

8. 顺序（等级）相关系数计算

设两个要素 x 和 y 有 n 对样本值，令R_1代表要素 x 的位次，R_2代表要素 y 的位次，$d_i{}^2=(R_{1i}-R_{2i})^2$ 代表要素 x 和 y 的同一组样本位次差的平方，则顺序（等级）相关系数的计算公式为：

$$r_s=1-\frac{6\sum d_i{}^2}{n(n^2-1)}$$

9. 秩相关系数的检验表

n	显著水平α		n	显著水平α	
	0.05	0.01		0.05	0.01
4	1.000	--	16	0.425	0.601
5	0.900	1.000	18	0.399	0.564
6	0.829	0.943	20	0.377	0.534
7	0.714	0.893	22	0.359	0.508
8	0.643	0.833	24	0.343	0.485
9	0.600	0.783	26	0.329	0.465
10	0.564	0.746	28	0.317	0.448
12	0.456	0.712	30	0.306	0.432
14	0.456	0.645	--	--	--

N 表示样本个数，α代表不同的置信水平，表中的数值r_α表示临界值，当$r_s>r_\alpha$时，可以认为两个要素是相关的。

10. 相关系数矩阵 ➔ 是一个对角线数值为 1 的对称矩阵

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

➢ 多要素间相关程度的测度

1. **偏相关与偏相关系数：**在多要素所构成的地理系统中，当我们研究某一个要素对另一个要素的影响或相关程度时，把其它要素的影响视为常数（保持不变），即暂不考虑其它要素的影响，而单独研究那两个要素之间的相互关系的密切程度时，则称为偏相关。用以度量偏相关程度的统计量，称为偏相关系数。

2. **偏相关系数的计算：**利用单相关系数来计算，假设有三个要素x_1, x_2, x_3，其两两之间的单相关系数（又称为零级相关系数）矩阵为：

$$R=\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}=\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

一级偏相关系数用形如$r_{12.3}$、$r_{13.2}$、$r_{23.1}$来表示。下标点后面的数字，代表在计算偏相关系数时保持的不变量，如$r_{12.3}$表示x_3保持不变，x_1和x_2之间的相关性。三个要素的一级偏相关系数共有 3 个，一级偏相关系数和系数矩阵分别为：

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-{r_{13}}^2)(1-{r_{23}}^2)}}\ ;\ r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-{r_{13}}^2)(1-{r_{23}}^2)}};$$

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-{r_{13}}^2)(1-{r_{23}}^2)}}\ ;\ R=\begin{bmatrix} 1 & r_{12.3} & r_{13.2} \\ r_{21.3} & 1 & r_{23.1} \\ r_{31.2} & r_{32.1} & 1 \end{bmatrix}$$

二级偏相关系数（假设有四个要素x_1, x_2, x_3, x_4）的计算公式，以$r_{12.34}$和$r_{13.24}$为例：

$$r_{12.34}=\frac{r_{12.3}-r_{14.3}r_{24.3}}{\sqrt{(1-{r_{14.3}}^2)(1-{r_{24.3}}^2)}}\ ;\ r_{13.24}=\frac{r_{13.2}-r_{14.2}r_{34.2}}{\sqrt{(1-{r_{14.2}}^2)(1-{r_{34.2}}^2)}}$$

3. **偏相关系数的性质**：

   1) 偏相关系数分布的范围为$[-1,1]$。

   2) 例如当$r_{12.3}$为正值时，表示x_3保持不变，x_1和x_2之间为正相关；当$r_{12.3}$为正值时，表示x_3保持不变，x_1和x_2之间为负相关。

   3) 偏相关系数的绝对值越大，表示其偏相关程度越大，例如当偏相关系数$r_{12.3}$的绝对值趋向于 1 表示x_3保持不变，x_1和x_2之间完全相关；当偏相关系数$r_{12.3}$的绝对值趋向于 0 表示x_3保持不变，x_1和x_2之间完全无关。

   4) 偏相关系数的绝对值不大于由同一系列资料所求得的复相关系数，即$r_{1.23} \geq r_{12.3}$

4. **偏相关系数的显著性检验：**偏相关系数的显著性检验，一般采用 t-检验法

t-检验法的统计量计算公式为：

$$t = \frac{r_{12.34\ldots m}}{\sqrt{1 - {r_{12.34\ldots m}}^2}} \times \sqrt{n - m - 1}$$

式中，$r_{12.34\ldots m}$为偏相关系数，n 为样本数，m 为自变量个数。

查 t 分布表，可得出不同显著水平上的临界值t_α：

若$t > t_\alpha$，则表示偏相关显著；反之，$t < t_\alpha$，则偏相关不显著。

5. **复相关系数：**反映几个要素与某一个要素之间的复相关程度，可以利用单相关系数和偏相关系数求得。若 y 为因变量，$x_1, x_2, x_3 \ldots x_k$为自变量，则将 y 和$x_1, x_2, x_3 \ldots x_k$之间的复相关系数记为$R_{y.12\ldots k}$，计算公式如下：

当有两个自变量时：$R_{y.12} = \sqrt{1 - (1 - {r_{y1}}^2)(1 - {r_{y2.1}}^2)}$

当有三个自变量时：$R_{y.123} = \sqrt{1 - (1 - {r_{y1}}^2)(1 - {r_{y2.1}}^2)(1 - {r_{y3.12}}^2)}$

当有 k 个自变量时：$R_{y.12\ldots k} = \sqrt{1 - \left(1 - {r_{y1}}^2\right)\left(1 - {r_{y2.1}}^2\right) \ldots \left(1 - {r_{yk.12\ldots(k-1)}}^2\right)}$

6. **复相关系数的性质**：

1) 复相关系数介于 0 到 1 之间。

2) 复相关系数越大，则表明要素（变量)之间的相关程度越密切。复相关系数为 1，表示完全相关；复相关系数为 0，表示完全无关。

3) 复相关系数不小于偏相关系数的绝对值。

7. **复相关系数的显著性检验（F-检验法)**，n 为样本数，k 为自变量个数，公式为：

$$F = \frac{{R_{y.12\ldots k}}^2}{1 - {R_{y.12\ldots k}}^2} \times \frac{n - k - 1}{k}$$

查 F-检验的临界值表可以得出不同显著水平上的临界值F_α：

1) 若$F > F_{0.01}$，表示复相关在置信度水平α=0.01 上显著，称为极显著；

2) 若$F_{0.05} < F \le F_{0.01}$，表示复相关在置信度水平$\alpha$=0.05 上显著；

3) 若$F_{0.1} \le F \le F_{0.05}$，表示复相关在置信度水平$\alpha$=0.1 上显著；

4) 若$F < F_{0.1}$，表示复相关不显著，即因变量 y 与 k 个自变量之间的关系不密切

➢ 空间相关性分析

1. **空间统计分析：**即空间数据的统计分析。

- 空间统计分析是以区域化变量理论为基础，以变异函数为主要工具，研究具有地

理空间信息特性的事物或现象的空间相互作用及变化规律的学科。

- 空间统计分析方法假设研究区中所有的值都是非独立的，相互之间存在相关性。
- 空间统计分析的核心在于认识与地理位置相关的数据间的空间依赖、空间关联或空间自相关，通过空间位置建立数据间的统计关系。

2. **空间统计分析的任务**：①揭示空间相关性；②利用相关规律进行未知点预测。
3. **区域化：**当一个变量呈空间分布时，称为区域化。
   - 区域变量具二重性：测量前是随机的，测量后是具体的。
4. **空间自相关：**通过相关分析可以检测两种现象（统计量）的变化是否存在相关性，若所分析的统计量为不同观察对象的同一属性变量，则称之为自相关。➔ 反映的是一个区域单元上的某种地理现象或某一属性值与邻近区域单元上同一现象或属性值的相关程度，是一种检测与量化从多个区域单元取样值变异的空间依赖性的空间统计方法
   - 当变量在空间上表现出一定的规律性，即不是随机分布则存在着空间自相关。
   - 空间自相关理论认为彼此之间距离越近的事物越相像。当某一测样点的属性值高，而其相邻点同一属性值也高，称为正相关；反之称为负相关。
   - 当空间自相关仅与两点间距离有关，称为各向同性；当考虑方向的影响时，可能在不同方向上属性值与距离的关系不同时，称为各向异性。
   - 半变异函/协方差函数云图能够检测已测样点间的空间自相关
5. **空间自相关的方法：**①空间权重矩阵；②全局空间自相关；③局部空间自相关
6. **空间权重矩阵：**定义一个二元对称空间权重矩阵 W，来表示 n 个位置的空间区域的邻近关系。式中W_{ij}表示区域 i 和 j 的临近关系，可以根据邻接标准或距离标准来度量。

$$W=\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix}$$

两种最常用的确定空间权重矩阵的规则：

1) 简单的二进制邻接矩阵：当区域 i 与 j 相邻接时$W_{ij}=1$，否则$W_{ij}=0$
2) 基于距离的二进制空间权重矩阵：当区域 i 与 j 的距离小于 d 时$W_{ij}=1$，否则$W_{ij}=0$

7. **全局空间自相关**：

两个用来度量空间自相关的全局指标：①Moran 指数、②Geary 系数

Moran 指数反映的是空间邻接或空间邻近的区域单元属性值的相似程度；

Geary 系数与 Moran 指数存在负相关关系

如果是位置（区域）的观测值，则该变量的**全局 Moran 指数 I 可以通过下式计算**：

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\times\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n}W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{S^2\sum_{i=1}^{n}\sum_{j\neq i}^{n}W_{ij}}$$

$$S^2=\frac{1}{n}\sum_{i}(x_i-\bar{x})^2\ ;\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

对于 Moran 指数，可以用**标准化统计量 Z** 来检验 n 个区域是否存在空间自相关关系，Z 的计算公式为：

$$Z=\frac{I-E(I)}{\sqrt{VAR(I)}}$$

➔ 当 Z 值为正且显著时，表明存在正的空间自相关，也就是说相似的观测值（高值或低值）趋于空间集聚；当 Z 值为负且显著时，表明存在负的空间自相关，相似的观测值趋于分散分布；当 Z 值为零时，观测值呈独立随机分布。

8. **局部空间自相关分析：**空间关联的局部指标（LISA）、G 统计量、Moran 散点图。

9. **空间关联的局部指标（Local indicators of spatial association）**

1) 每个区域单元的 LISA，是描述该区域单元周围显著的相似值区域单元之间空间集聚程度的指标。

2) 所有区域单元 LISA 的总和与全局的空间关联指标成比例。

LISA 包括局部 Moran 指数（Local Moran）和局部 Geary 指数（Local Geary）

局部 Moran 指数的定义为：

$$I_i=\frac{(x_i-\bar{x})}{S^2}\sum_{j}W_{ij}(x_j-\bar{x})=\frac{n(x_i-\bar{x})\sum_j W_{ij}(x_j-\bar{x})}{\sum_i(x_i-\bar{x})^2}$$

$$=\frac{nZ_i\times\sum_j W_{ij}Z_j}{Z^TZ}=Z_i'\sum_j W_{ij}Z_j'$$

其中Z_i'和Z_j'是经过标准差标准化的观测值。

局部 Moran 指数检验的标准化统计量为：$Z(I_i)=\frac{I_i-E(I_i)}{\sqrt{VAR(I_i)}}$

10. **G 统计量**

全局 G 统计量的计算公式：$G=\frac{\sum_i\sum_j W_{ij}x_ix_j}{\sum_i\sum_j x_ix_j}$

每个区域单元的 G 统计量计算公式为：$G_i=\frac{\sum_i W_{ij}x_j}{\sum_j x_j}$

G 统计量的检验值为：$Z(G_i)=\frac{G_i-E(G_i)}{\sqrt{VAR(G_i)}}$

➔ 当值为正且显著时，表示在该区域单元周围，高观测值的区域单元趋于空间集聚；当值为负且显著时表示低观测值的区域单元趋于空间集聚。

- 与 Moran 指数只能发现相似值(正关联)或非相似性观测值（负关联）的空间集聚模式相比，G 统计量能够探测出区域单元属于高值集聚还是低值集聚的空间分布模式。

11. **Moran 散点图**：以(W_Z, Z)为坐标点的散点图

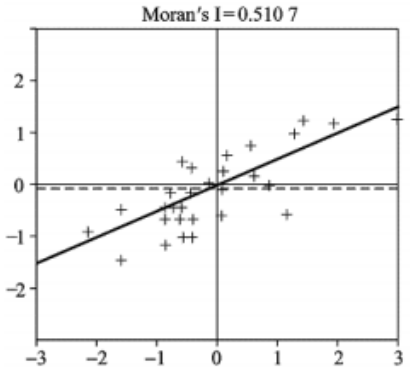
➔研究局部的空间不稳定性，对空间滞后因子W_Z和 Z 数据对进行了可视化的二维图示

➔全局 Moran 指数，可以看作是W_Z对于 Z 的线性回归系数，对界外值以及对 Moran 指数具有强烈影响的区域单元，可通过标准回归来诊断出。

- **Moran 散点图的四个象限**

分别对应于区域单元与其邻居之间四种类型的局部空间关联形式

1) 第一象限代表了高观测值的区域单元被同是高值的区域所包围的空间关联形式；
2) 第二象限代表了低观测值的区域单元被高值的区域所包围的空间关联形式；
3) 第三象限代表了低观测值的区域单元被同是低值的区域所包围的空间关联形式；
4) 第四象限代表了高观测值的区域单元被低值的区域所包围的空间关联形式。



第九章・空间点模式分析

➢ 空间点模式分析概述

1. **空间点模式分析（Point Pattern Analysis，PPA）**：根据实体或事件的空间位置研究其分布模式的方法称为空间点模式分析，PPA 是一类重要的空间分析方法。
2. **空间点模式**：是研究区域 R 内的一系列点的组合：$S_1=(x_1,y_1),\ldots,S_i=(x_i,y_i)$

➔ S_1是第 i 个观测时间的空间位置

3. **点模式的三种基本类型：**聚集分布、随机分布、均匀分布

4. **两种点模式分析方法**：

   1) 以聚集性为基础的基于密度的方法，主要有样方分析法和核函数方法两种

   2) 以分散性为基础的基于距离的技术，通过测度最近邻点的距离分析点的空间分布模式的方法，主要包括最邻近指数、G-函数、F-函数、K-函数方法等。

5. **样方分析法（quadrat analysis，QA）**：属于基于密度的点模式分析方法，通过空间上点分布密度的变化探索空间分布模式，一般使用随机分布模式作为理论上的标准分布，将 QA 计算的点密度和理论分布做比较，判断点模式属于聚集分布、均匀分布还是随机分布。

   - **一般过程：**①将研究区域划分为规则的正方形网格区域 ➔ ②统计落入每个网格中点的数量（由于点在空间上分布的疏密性，有的网格中点的数量多，有的网格中点的数量少，有的网格中点的数量甚至为零） ➔ ③统计出包含不同数量点的网格数量的频率分布 ➔ ④将观测得到的频率分布和已知的频率分布或理论上的随机分布（如泊松分布）作比较，判断点模式的类型。
   - **QA 方法对分布模式判别产生影响的主要因素：**①样方的形状、②采样的方式、③样方的起点、方向和大小。【无论采用何种形式的样方，要求网格形状和大小必须一致，避免在空间上的采样不均匀】
   - **最优的样方尺寸：**根据区域的面积和分布于其中的点的数量确定，计算公式为：

$$Q = \frac{2A}{n}$$

     Q 是样方的尺寸（面积），A 为研究区域的面积，n 为研究区域中点的数量
   - **样方分析中点模式的显著性检验：**通过实际的分布观测频数与某种分布模式（均匀分布、随机分布、聚集分布）进行比较，通过定量化地计算频率分布的差异进行判断。

     常用的检测方法：①根据频率分布比较的 K-S 检验、②根据方差均值比的x^2检验

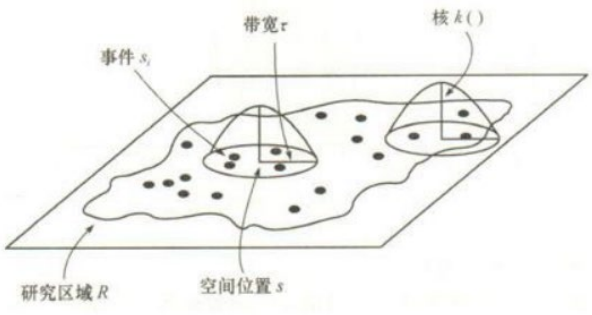
6. **核密度估计法（kernel density estimation, KDE）：**属于基于密度的点模式分析方法，认为地理事件可以发生在空间的任何位置上，但是不同位置上事件发生的概率不一样

   核密度估计法的原理：使用事件的空间密度分析表示空间点模式，点密集的区域事件发生的概率高，点稀疏的区域事件发生的概率低。

设在研究区域 R 内分布有 n 个事件 S，S 处的点密度估计为：

$$\hat{\lambda}_\tau(S) = \sum_{i=1}^{n} \frac{1}{\tau^2} k(\frac{s - s_i}{\tau})$$

$k()$表示核的权重函数；$\tau > 0$是带宽（以 s 为源点的曲面在空间上延展的宽度，会影响分布密度估计的光滑程度）；$(s - s_i)$是需要进行密度估值的点 s 和s_i之间的距离。



影响 KDE 的主要因素是：$k()$函数的形式和带宽

核密度估计法的特点：①核函数$k()$的值在$d_{ij} = 0$时最大，随着距离d_{ij}的增加，k()值减小；②在空间上点 s 处的密度估计值是已知事件对于该点的综合影响；③核函数中的带宽确定了事件的影响范围。

带宽值的选择是有弹性的：需要根据不同的带宽值进行试验，探索估计的点密度曲面的光滑程度，以检验带宽尺度变化的影响。可以根据 R 中点的位置调整带宽的值，这种带宽的局部调节是自适应的方法，在事件密集的子区域内，具有更加详细的密度变化信息，因此带宽值取小一点。

7. **最邻近距离法：**属于基于距离的点模式分析方法，使用最邻近的点对之间的距离描述分布模式，形式上相当于密度的倒数（每个点代表的面积）表示点间距。

一般过程：计算最邻近的点对之间的平均距离➔比较观测模式和已知模式之间的相似性➔一般将随机模式作为比较的标准➔如果观测模式的最邻近距离大于随机分布的最邻近距离，则观测模式趋向于均匀分布；如果观测模式的最邻近距离小于随机分布模式的最邻近距离，则趋向于聚集分布。

最邻近距离：指任意一个点到其邻近的点对之间的距离。利用欧式距离公式，得到研究区域中每个事件的最邻近点及其距离。将点s_i的最邻近距离记为$d_{min}(s_i)$。

- 点对之间的最邻近距离不一定是相互的，在 CSR 模式（完全随机模式）中，超过 60%的最邻近是相互的邻近。

8. **最邻近指数测度方法（NNI: Nearest Neighbor Index）**：对于同一组数据，在不同的分布模式下得到 NNI 是不同的，根据观测模式的 NNI 计算结果与 CSR 模式下的 NNI 比较，就可判断分布模式的类型。

一般过程：

1) 对研究区内的任意一点都计算最邻近距离

2) 取这些最邻近距离的均值作为评价模式分布的指标$\bar{d}_{min} = \frac{1}{n}\sum_{i=1}^{n} d_{min}(s_i)$

3) 计算 CSR 模式中的平均的最邻近距离的期望值$E(d_{min})$

4) 通过最邻近距离的计算和比较就可以评价和判断分布模式：

两种方法：基于最邻近距离或基于最邻近指数 R（$R = \frac{\bar{d}_{min}}{E(d_{min})} = 2\bar{d}_{min}\sqrt{\frac{n}{A}}$）

a) 若$R = 1$，说明观测事件过程来自于完全随机模式 CSR，属于随机分布；

b) 若$R < 1$，说明大量事件点在空间上相互接近，属于空间聚集模式；

c) 若$R > 1$，说明点之间的最邻近距离大于 CSR 过程的最邻近距离，时间模式中的空间点是相互排斥的，趋于均匀分布模式。

d) 特殊情况：①极端聚集（所有事件发生在研究区域的同一位置上，R=0）②极端均匀（均质区域上邻近的 3 个点构成等边三角形，空间被正六边形划分，点位于正六边形的中心，实质是克里斯泰勒中心分布的模式，R=2.149）

显著性检验的过程：

1) 首先计算观测的平均最邻近距离和 CSR 的期望平均距离的差异：$\bar{d}_{min} - E(d_{min})$

2) 利用这一差异和其标准差 $SE_r = \sqrt{VAR(\bar{d}_{min} - E(d_{min}))}$ 作比较：

a) 计算的差异与其标准差比较相对较小，说明这种差异在统计上不显著，即点模式属于 CSR

b) 计算的差异与其标准差比较相对较大，说明这种差异在统计上显著，即点模式不属于 CSR

3) 构造一个服从正态分布的统计量 Z：$Z = \frac{\bar{d}_{min} - E(d_{min})}{SE_r}$

当显著性水平为 a 时，Z 的置信区间为$-Z_a \le Z \le Z_a$

若$Z < -Z_a$或$Z > Z_a$，那么观测模式和 CSR 之间存在显著差异

若 Z 的符号为负，则模式趋于聚集；若 Z 的符号为正，则模式趋于均匀

➢ G 函数与 F 函数

1. **G 函数：**使用所有的最近邻事件的距离构造出一个最邻近距离的累积频率函数

$$G(d) = \frac{\#(d_{min}(s_i) \le d)}{n}$$

s_i是研究区域的一个事件，n 是事件的数量，d 是距离

$\#(d_{min}(s_i) \le d)$是距离小于 d 的最邻近点的计数

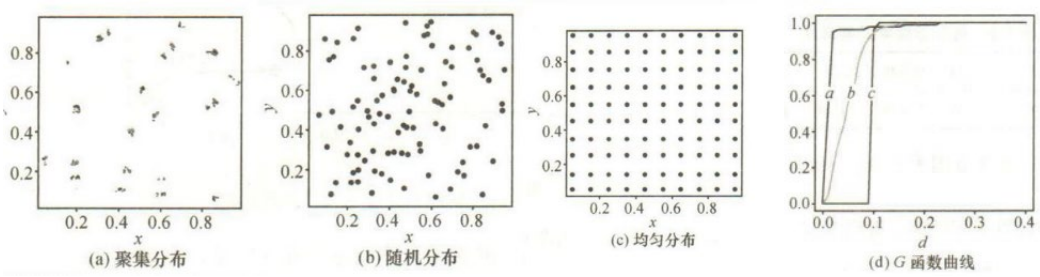
随着距离 d 的增大，最邻近距离点累积个数会增加，$G(d)$也相应增大，因此$G(d)$为累积分布，直到 d 等于最大的最邻近距离，这时最邻近距离点个数最多，$G(d)$的值为 1

计算$G(d)$的一般过程:

1) 计算任意一点到其最邻近点的距离d_{min}
2) 将所有的最邻近距离列表，并按照大小排序
3) 计算最邻近距离的变程 R 和组距 D，其中，$R = max(d_{max}) - max(d_{min})$，组距 D 可按照 Sturges 规则（分类的数量 x 介于$2^n < x < 2^{n+1}$）确定
4) 根据组距上限值，累积计数点的数量，并计算累积频数$G(d)$
5) 画出关于 d 的曲线图。

用$G(d)$曲线的形状分析空间点模式:

1) 如果$G(d)$在短距离内迅速增长，表明点空间分布属于聚集模式
2) 如果$G(d)$先缓慢增长后迅速增长，表明点空间分布属于均匀模式



2. **F 函数：**与 G 函数类似，使用最邻近距离的累积频率分布描述空间点模式类型的一阶邻近测度方法，F 函数记为$F(d)$。区别于 G 函数，F 函数首先在被研究的区域中产生一新的随机点集$P(P_1, P_2 \dots P_i \dots P_m)$，其中$P_i$是第 i 个随机点的位置。然后计算随机点到事件点 S 之间的最邻近距离，再沿用 G 函数的思想，计算不同最邻近距离上的累积点数和累积频率。

$$F(d) = \frac{\#(d_{min}(P_i, S) \le d)}{n}$$

$d_{min}(P_i, S)$表示从随机选择的 pi 点到其最邻近的事件点 S 的最邻近距离

基于 F 函数的分布模式判别：

1) 在 F 函数中，若 F 函数曲线缓慢增加到最大，则表明是聚集模式；

2) 若 F 函数曲线快速增加到最大，则表明是均匀分布模式。

3. F 函数与 G 函数的比较

1) **共同点：** F 函数和 G 函数都采用了最邻近距离的思想描述空间点模式

2) **二者的本质差别**：G 函数主要是通过事件之间的邻近性描述分布模式，F 函数则主要通过选择的随机点和事件之间的分散程度来描述分布模式。F 函数曲线和 G 函数曲线呈相反的关系。

➢ K 函数与 L 函数

1. 为了研究地理事件空间依赖性与尺度的关系➔ Ripley 提出了基于二阶性质的 K 函数方法；Bessage 将 K 函数变换为 L 函数➔ K 函数和 L 函数是描述在各向同性或均质条件下点过程空间结构的良好指标。

2. **K 函数：** $K(d)$定义为以任意点为中心，半径 d 范围内点的数量的期望除以点密度λ

$$K(d) = \frac{E\left(\#S \in (S_i, d)\right)}{\lambda}$$

K 函数的估计：（其中 a 是研究区域的面积，n 是研究区域内点的数量）

$$\hat{K}(d) = \frac{1}{n\lambda}\sum_{1}^{n} \#S \in (S_i, d) = \frac{a}{n^2}\sum_{1}^{n} \#S \in (S_i, d)$$

K 函数的边缘效应与校正： 当d_{ij}超出研究区域的范围时，需要进行校正以消除边缘效应，引入校正因子w_{ij}。周长比例校正法和面积比例校正法最常用，其中面积比例校正法的w_{ij}等于以事件S_i为中心，以S_iS_j为半径的圆域在研究区域内部的面积占该圆域的比例。

CSR 模式下，$K(d) = \pi d^2$ ➔

①$K(d) = \pi d^2$➔随机模式；②$K(d) > \pi d^2$➔聚集模式；③$K(d) < \pi d^2$➔均匀模式

3. **L 函数：** 以零为比较标准的规格化函数

$$L(d) = \sqrt{\frac{K(d)}{\pi} - d}$$

L 函数的估计：

$$\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi} - d}$$

在 CSR 模式下，$L(d) = 0$ ➔在 L 函数图中：

1) 正的峰值表示点在这一尺度上的聚集或吸引；

2) 负的峰值表示点的均匀分布或空间上的排斥

- 观测模式随着尺度 d 的变化而变化。在小尺度上表现出一阶方法所揭示的均匀性；在较大尺度上表现出的是聚集性。

第十章 · 地统计分析

➢ 地统计分析概述

1. **地统计分析的由来：**20 世纪 50 年代，南非采矿工程师 Daniel Krige 总结多年金矿勘探经验，提出了克里金估计方法(kriging)的雏形 ➔ 20 世纪 60 年代初期，法国地质数学家 Georges Matheron 提出数学形式的区域化变量，严格地给出了基本变异函数的定义和一般克里金估计方法。
2. **Daniel Krige 提出的地统计分析的基本思路**：根据样品点的空间位置和样品点之间空间相关程度的不同，对每个样品观测值赋予一定的权重，进行移动加权平均，估计被样品点包围的未知点的矿产储量。
3. **地统计学（Geostatistics）**：也称为地质统计学，是一门以区域化变量理论为基础，以变异函数为主要工具，研究那些分布于空间上既有随机性又有结构性的自然或社会现象的科学。
4. **地统计学的特点**：
   1) **样本点的空间相关性**：传统统计中不同样本点仅具有随机性，样本点之间保持空间独立性。然而，地统计中样本点不仅具有随机性，同时样本点之间具有空间相关性。
   2) **一次性样本采集**：传统统计分析同一空间位置处可以多次采样数据。实际地统计分析中，样本区域中每一个空间位置多为一次采样数据。

➢ 区域化变量理论

1. **区域：**当空间被赋予地学含义时，地学工作者习惯称其为区域。
   - 地理学研究的基本任务之一：发现地表空间的区域差异
2. **区域化：**当一个专题变量分布于空间，呈现一定的结构性和随机性时，在地统计学上

称之为区域化。

3. **区域化变量(regionalized variable)**：是区域化随机变量的简称，是空间位置相关的随机变量，为具有内在空间结构的随机变量，是随机场的简化。用$Z(x)$表示在空间位置 x 处的专题变量取值。

   - 在任意点 x 处，$Z(x)$是一个随机变量 ➔随机性
   - 点x_1和x_2处的随机变量$Z(x_1)$和$Z(x_2)$通常是不独立的 ➔结构性

4. **区域化变量理论**：揭示区域化变量空间结构和统计性质的理论。重点研究区域化随机变量的各种空间结构和统计性质。

➢ 空间变异函数

1. **变异函数**：变异函数是描述区域化随机变量空间结构的有效数学工具，为两个随机变量$Z(x)$和$Z(x+h)$之间增量的方差的一半。

   在二阶平稳性或内蕴平稳性假设下，原始变异函数$\gamma_x(h)$规约为单纯的空间滞后 h 的$r(h)$，与空间位置 x 无关，即：

$$\gamma(h)=\frac{1}{2}Var[Z(x+h)-Z(x)]=\frac{1}{2}E\{[Z(x+h)-Z(x)]-E[Z(x+h)-Z(x)]\}^2$$

$$=\frac{1}{2}E[Z(x+h)-Z(x)]^2=C(0)-C(h)$$

2. **平稳性假设**：地统计中的数据多为区域中每个空间位置的一次采样数据。

   为了满足总体规律推断中多个样本（大样本）的数据要求，地统计中使用平稳(second-order stationary)或内蕴(intrinsic stationary)假设下多个空间位置采样数据（每个位置一次采样数据）来替代单个位置上的多次采样数据（传统统计的采样数据）。

3. **严格平稳性**：存在 n 个随机变量的联合分布$F\big(Z(x_1),Z(x_2),\dots,Z(x_n)\big)$，严格平稳性指随机变量联合分布的空间位移不变性。

$$F\big(Z(x_1),Z(x_2),\dots,Z(x_n)\big)=F\big(Z(x_1+h),Z(x_2+h),\dots,Z(x_n+h)\big)$$

4. **弱平稳性假设**：满足位移不变的联合概率分布的区域化随机变量较少见，严格平稳性的验证非常困难 ➔ 容易满足和验证的是分布参数(矩)的平稳性，即弱平稳性假设。

   常用的弱平稳性假设：二阶平稳性和内蕴性假设

5. **二阶平稳性假设**：区域化变量$Z(x)$满足下列两个条件，则称其满足二阶平稳性假设。

   1) 在研究范围内，区域化变量$Z(x)$的期望存在且为常数，即$E[Z(x)]=m$

   2) 在研究范围内，区域化变量$Z(x)$的协方差函数存在且为空间滞后 h 的函数，与空

间位置 x 无关，即：

$$Cov[Z(x),Z(x+h)] = E\{[Z(x+h)-m][Z(x)-m]\} = E[Z(x+h)Z(x)] - m^2 = C(h)$$

当$h=0$时，方差函数存在且为常数，即：

$$Var[Z(x)] = Cov[Z(x),Z(x)] = E[Z(x)-m]^2 = C(0)$$

6. **内蕴性假设**：区域化变量 Z(x)满足下列两个条件，则称其满足内蕴性假设。

   1) 在研究范围内，区域化变量$Z(x)$的增量的期望为 0，即$E[Z(x+h)-Z(x)]=0$

   2) 在研究范围内，区域化变量$Z(x)$的增量的方差存在且为空间滞后 h 的函数，与空间位置 x 无关，即：

$$Var[Z(x+h)-Z(x)] = E\{[Z(x+h)-Z(x)] - E[Z(x+h)-Z(x)]\}^2$$

$$= E[Z(x+h)-Z(x)]^2 = 2\gamma(h)$$

   - 二阶平稳性是比内蕴性更严格的弱平稳性假设

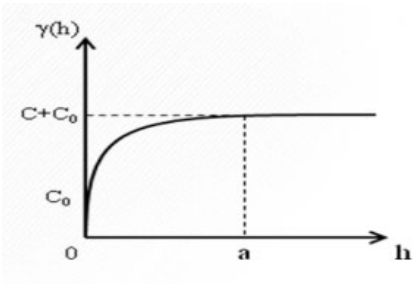
7. **变异函数模型拟合/变异函数曲线（variography）**：变异函数值随着空间滞后 h 的增大而单调增加，达到极限值后上下波动。

   变异函数$\gamma(h)$具有三个参数$\{a, C_0, C+C_0\}$

   1) **块金值C_0**：是空间滞后为 0 时的变异函数值，是测量误差和低于采样间距的随机变异的综合反映。

   2) **基台值$C+C_0$**：当空间滞后 h 超过变程 a 时，变异函数$\gamma(h)$在一个极限值$\gamma(\infty)$附近摆动，这个极限值称为基台值。

   3) **变程 a**：是变异函数达到基台值$C+C_0$时的空间滞后 h，反映了空间自相关的最大距离。



变异函数曲线
(variography)

8. 区域化变量的取值 Z 由大尺度趋势μ、微尺度空间相关变异γ和纯随机变异ε三部分构成，即$Z=\mu+\gamma+\varepsilon$

   1) μ：期望，是一种趋势表示。

   2) γ：微尺度空间相关变异，是去除趋势后具有内在空间自相关性的残余值。

   3) ε：纯随机变异，是不存在空间自相关性的独立噪声(如测量误差)

➔ 测量误差和采样间距(采样尺度)以下的微尺度空间相关残余值一起构成块金值C_0，采样间距(采样尺度)以上的微尺度空间相关残余值的变异函数值为 C。

9. **理论变异函数模型**:

1) 有基台值模型:

球状模型、指数模型、高斯模型、有基台线性模型和纯块金效应模型等。

2) 无基台值模型：幂函数模型、无基台线性模型和对数模型等。

3) 孔穴效应模型。

(1)球状变异函数模型

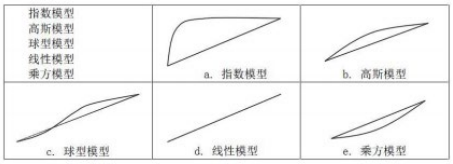
$$\gamma(h)=\begin{cases}0, & h=0\\ C_0+C_1[1.5(h/a)-0.5(h/a)^3], & 0\le h\ge a\\ C_0+C_1, & h>a\end{cases}$$

(2)指数变异函数模型

$$\gamma(h)=\begin{cases}0, & h=0\\ C_0+C_1(1-e^{-h/a}), & h<0\end{cases}$$

(3)高斯变异函数模型

$$\gamma(h)=\begin{cases}0, & h=0\\ C_0+C_1[1-e^{-(h/a)^2}], & h>0\end{cases}$$



(4)有基台线性变异函数模型

$$\gamma(h)=\begin{cases}C_0, & h=0\\ Ah, & 0<h\le a\\ C_0+C_1, & h>a\end{cases}$$

(5)纯块金变异函数模型

$$\gamma(h)=\begin{cases}0, & h=0\\ C_0, & h>0\end{cases}$$

(6)幂函数变异函数模型

$$\gamma(h)=C_0+C_1h^{\lambda},\ 0<\lambda<2$$

(7)无基台线性变异函数模型

$$\gamma(h)=\begin{cases}C_0, & h=0\\ Ah, & h>0\end{cases}$$

(8)对数变异函数模型

$$\gamma(h)=\log h$$

(9)孔穴效应变异函数模型

$$\gamma(h)=C_0+C[1-e^{-\frac{h}{a}}\cos(2\pi\frac{h}{b})]$$

➢ 克里金估计方法

1. **估计在时间空间域的分类**:

1) **在时间域**：服务于不同目的估计分别称为滤波(除去噪音)、平滑(找出趋势)和预测(计算未来值)。

2) **在空间域**：估计分为内插(计算研究区域内的未知值) 和外推(计算研究区域外的未知值，又称为预测)。

2. **克里金估计**：利用区域化变量的结构性质进行估值应用

- 克里金插值和克里金预测统称为克里金估计。

3. **克里金估计方法的分类**:

1) **点估值（最基本的方法）包括**：普通克里金估计、简单克里金估计、泛克里金估计、协同克里金估计、对数克里金估计、指示克里金估计、析取克里金估计和概率克里金估计、普通协同克里金估计、协同泛克里金估计和协同指示克里金估计等方法。

2) **块段估值**：块段的值可以通过赋予块段平均值给块段中心点来转化为点估值；或者把块段离散为若干点的集合，从而转化为点估值。

4. **普通克里金估计：**在内蕴假设(或二阶平稳假设)下期望未知的区域化变量估值方法

➔ $Z(x)=m+Y(x)$，期望 m 未知，残余$Y(x)$的期望为零。

普通克里金估计方法的估计公式：$Z^*(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i)$

式中，$Z^*(x_0)$是待估位置x_0的估值，$Z(x_i)$是已知位置x_i的观测值，λ_i是分配给$Z(x_i)$的权重，n 是估计使用的观测值个数。

普通克里金估计方法的无偏估计条件：$E[Z^*(x_0)-Z(x_0)]=0 \Leftrightarrow \sum_{i=1}^{n}\lambda_i=1$

普通克里金估计方法的最优估计条件：估计误差方差最小

➔ 引入拉格朗日乘数-2μ，将条件（无偏估计条件）极值（估计方差最小）问题化为无条件表达式的极值问题求解，获得普通克里金方程组（如下），求解出权重系数，代入普通克里金估计方法估计公式进行代估点的估值。

$$\begin{cases}\sum_{j=1}^{n}\lambda_j Cov[Z(x_i),Z(x_j)]-\mu=Cov[Z(x_i),Z(x_0)],i=1,2,\dots,n\\ \sum_{i=1}^{n}\lambda_i=1\end{cases}$$

5. **泛克里金估计：**区域化变量非平稳（存在漂移）的克里金估计

➔ $Z(x)=m(x)+Y(x)$，期望$m(x)$代表趋势项/漂移；残余$Y(x)$具有内蕴性（或二阶平稳性）且期望为零。

泛克里金估计方法的估计公式为：$Z^*(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i)$

式中，$Z^*(x_0)$是待估位置x_0的估值，$Z(x_i)$是已知位置x_i的观测值，λ_i是分配给$Z(x_i)$的权重，n 是估计使用的观测值个数。

泛克里金估计方法的无偏估计条件：

$$E[Z^*(x_0)-Z(x_0)]=0 \Leftrightarrow \sum_{i=1}^{n}\lambda_i f_l(x_i)=f_l(x_0),l=0,1,\dots,k$$

普通克里金估计方法的最优估计条件：估计误差方差最小

➔ 引入拉格朗日乘数-2μ，将条件（无偏估计条件）极值（估计方差最小）问题化为无条件表达式的极值问题求解，获得泛克里金方程组（如下），求解出权重系数，代入普通克里金估计方法估计公式进行代估点的估值。

$$\begin{cases}\sum_{j=1}^{n}\lambda_j Cov[Z(x_i),Z(x_j)]-\sum_{l=0}^{k}\mu_l f_l(x_i)=Cov[Z(x_i),Z(x_0)],i=1,2,\dots,n\\ \sum_{i=1}^{n}\lambda_i f_l(x_i)=f_l(x_0),l=0,1,\dots,k\end{cases}$$

6. **协同克里金估计**：多个变量的构成的协同区域化变量

协同克里金估计的二阶平稳性假设（以两个变量为例）：

1) 每一个变量的期望存在且为常数，$E[Z_k(x)] = m_k, k = 1,2$

2) 每一个变量的空间协方差为空间滞后 h 的函数，与绝对空间位置无关，

$$Cov[Z_k(x+h), Z_k(x)] = C_{kk}(h), k = 1,2$$

3) 两个变量的交叉协方差函数为空间滞后 h 的函数，与绝对空间位置无关，

$$Cov[Z_k(x), Z_{k'}(x+h)] = C_{kk'}(h), \quad k, k' = 1,2$$

- 交叉协方差中k, k'的顺序不能颠倒。

➔ 假设区域化变量$Z_2(x)$为主变量，观测值的个数为n_2，区域化变量$Z_1(x)$为辅助变量，观测值的个数为n_1。$Z_2(x)$比$Z_1(x)$难于观测，$n_2 < n_1$。综合利用$Z_1(x)$和$Z_2(x)$的观测值对$Z_2(x)$在x_0处进行估计。

协同克里金估计方法的估计公式为：

$$Z_2^*(x_0) = \sum_{i=1}^{n_1} \lambda_{1i} Z_1(x_{1i}) + \sum_{j=1}^{n_2} \lambda_{2i} Z_2(x_{2j})$$

式中，$Z^*(x_0)$是待估位置x_0的估值，$Z(x_i)$是已知位置x_i的观测值，λ_i是分配给$Z(x_i)$的权重，n 是估计使用的观测值个数。

协同克里金估计方法的无偏估计条件：

$$\begin{cases} \sum_{i=1}^{n_1} \lambda_{1i} = 0 \\ \sum_{j=1}^{n_2} \lambda_{2j} = 1 \end{cases}$$

7. **指示克里金估计**：限制特异值的影响，适应分布未知的情形，适用于非连续取值(包括名义数据)的非参数估计的指示克里金估计方法。

设有区域化变量$Z(x)$，通过如下**指示函数**将其转化为指示变量，取值为{0,1}。

$$I(x; Z_k) = \begin{cases} 1, Z(x) \le Z_k \\ 0, Z(x) > Z_k \end{cases}, Z_k \text{为阈值}$$

指示变量的内蕴性假设为：

1) 不同空间位置的两指示变量的增量的期望为零：$E[I(x+h; Z_k) - I(x; Z_k)] = 0$

2) 指示变异函数仅为空间滞后 h 的函数，与空间位置无关，其中$\gamma(h; Z_k)$表示指示变异函数：

$$E[I(x+h; Z_k) - I(x; Z_k)]^2 = 2\gamma(h; Z_k)$$

指示克里金估计方法的估计公式：

$$I^*(x_0;Z_k)=\sum_{i=1}^{n}\lambda_i Z(k)\,I(x_i;Z_k)$$

式中，$I^*(x_0;Z_k)$是在待估位置x_0的指示变量估计值，$I(x_i;Z_k)$是位置x_i观测值的指示化，$\lambda_i Z(k)$是分配给$I(x_i;Z_k)$的权重，n 是估计使用的观测值个数

求解权重需要满足无偏性条件和估计误差方差最小两个条件，即：

$$\begin{cases} E[I(x;Z_k)^*-I(x;Z_k)]=0 \\ Var[I(x;Z_k)^*-I(x;Z_k)]=min \end{cases}$$

8. **其他克里金估计方法类型的适用范围**：
   1) **对数克里金估计：**变量服从对数正态分布
   2) **普通协同克里金估计、协同泛克里金估计和协同指示克里金估计**：综合多个角度，全面利用区域化变量的结构性质，对单个特性建模的克里金估计进行组合。
9. **克里金估计模型的有效性检验**：
   1) 选择部分数据作为构造变异函数和克里金估计模型的训练数据，选择另外部分数据作为模型有效性的检验数据。
   2) 使用全部样本数据作为检验数据，如交叉验证使用的检验数据。交叉验证方法比较了所有点的测量值和估计值。
10. **采样设计准则**：
    1) 尽可能增加样本点个数；
    2) 尽可能采集靠近估计点的样本点；
    3) 尽可能使样本点之间彼此远离。
11. **克里金估计方法**对待估点的估值综合利用了待估点自身的结构信息(如方差和期望)、样本点和待估点之间的结构信息（如样本点到待估点的平均距离）和相关样本点内部结构信息（样本点之间的协方差等），它比一般简单距离加权平均方法具有更高的估计精度。
12. **克里金估计误差方差**同样综合利用了各种结构信息，评价不同待估位置估值的不确定性，它比简单统计指标（如遥感影像分类结果的混淆矩阵中的各种精度指标）对不确定性的评价更加精细。

第十一章 · 地理加权回归分析技术

➢ 地理加权回归分析技术

1. **空间异质性（spatial heterogeneity）**：是指生态、社会等空间过程和格局在空间分布上的不均匀性及其复杂性。
2. **地理加权回归分析（Geographically Weighted Regression，GWR）**
3. **回归分析**：是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法
4. **基本地理加权回归分析模型**：

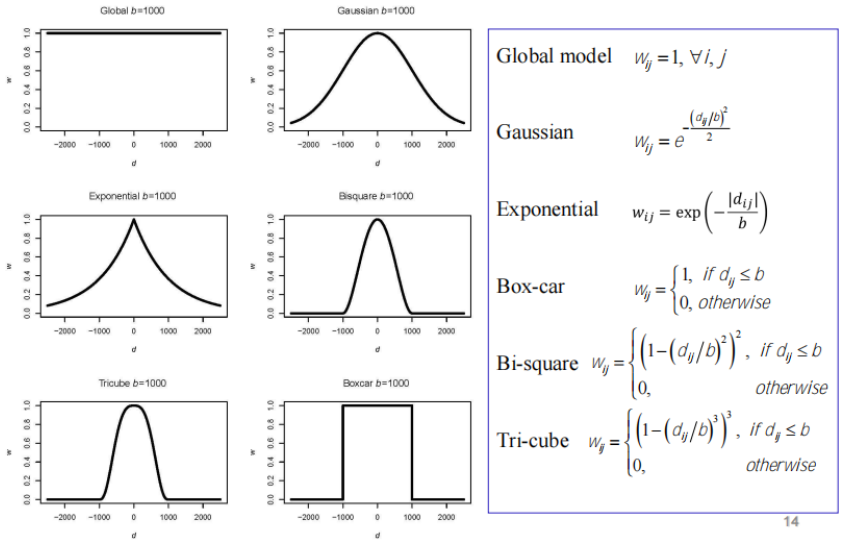
$$y_i = \beta_0(u_i, v_i) + \sum_{k=1,n} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

解算算子（加权线性最小二乘）：

$$\hat{\beta}_i = (X^T w_i X)^{-1} X^T w_i y$$

- 基于距离的权重计算：距离越近，权重越高；距离越远，权重越低。

5. **权重核函数**：



6. **带宽（Bandwidth, b）**：
   1) **固定型带宽（Fixed weighting scheme）**：定义一个固定的距离阈值 b，将固定值 b 用于所有点的权重计算
   2) **可变型带宽（Adaptive weighting scheme）**：定义一个正整数 N，针对任意求解位置点，计算数据点到该点的距离，取距其第 N 近的距离值作为当前的带宽值 b

➔ 不同类型的带宽带来不同的效果 ➔ 空间数据均匀分布/不均匀分布

➢ 多尺度地理加权回归分析技术

1. **Mixed GWR (Brunsdon et al., 1999)**：假设仅有部分参数估计在研究区域内是变化的，而其余的是不变的

$$y_i = \sum_{j=1,k_a} a_j x_{ij}(a) + \sum_{j=1,k_b} b_l(u_i, v_i) x_{il}(b) + \varepsilon_i$$

2. **GWR with flexible bandwidths（FBGWR）**：假设针对不同的参数采用变化的带宽对模型进行估计

$$y_i = \beta_{bw_0}(u_i, v_i) + \sum_{k=1,n} \beta_{bw_k}(u_i, v_i) x_{ik} + \varepsilon_i$$

3. **PSDM GWR**：针对不同的参数采用不同的距离度量和带宽进行估计

$$y_i = \beta_{DM_0,bw_0}(u_i, v_i) + \sum_{k=1,n} \beta_{DM_k,bw_k}(u_i, v_i) x_{ik} + \varepsilon_i$$

第十二章 · 空间分析建模及工作流技术

➢ 空间分析建模

1. 空间分析建模的概念模式 / 模型的概念化的基本步骤

1) **明确问题**：分析问题的实际背景，弄清建立模型的目的，掌握所分析对象的各种信息，不仅要明确所要解决的问题是什么，要达到什么样的目标，还要明确实际问题的解决途径和所需要的数据。

2) **分解问题**：找出实际问题有关的因素，通过假设把所研究的问题进行分解、简化，明确模型中需要考虑的因素以及它们在过程中的作用，并准备有关的数据集。

3) **收集数据**：确定并收集所需要的数据，包括一些相关的信息、关系、模式等。

4) **实施空间分析**：运用数学知识和空间分析工具描述问题中变量间的关系。

5) **结果检验**：运行所得到的模型、解释模型的结果或把运行结果与实际观测对比。如果与实际状况符合或基本一致，表明模型是符合实际问题的；否则，需要修改模型或其参数，重复前面的建模过程，直到模型的结果满意为止。

6) **应用分析结果**：在对模型满意的前提下，可以运用模型得到对结果的分析，产生相应的建议、对策、方案等。

2. **空间分析建模的方法（根据 GIS 空间分析建模的目的分类）**：

   1) **以特征为主的描述模型(descriptive model)**：是一类用描述方法研究区域中的实体类型、特征、相互之间的空间关系和实体属性特征的模型。➔ 回答“是什么”的地理问题，或者描述某类形象存在的环境条件。

   2) **提供辅助决策信息和解决方案为目的的过程模型(process model)**：运用数学分析方法建立表达式，模拟地理现象的形成过程的模型称为空间分析的过程模型，也叫空间分析的处理模型。➔ 回答“应当如何”之类的地理问题

   3) **面向应用的应用模型**：指空间分析的理论、方法和过程在与空间数据处理有关的领域的具体应用。

   - **根据使用的方法分类**：①随机模型、②确定性模型
   - **根据逻辑分类**：①归纳模型、②演绎模型

3. **空间分析功能与各种领域专用模型的结合**：①基于 GIS 二次开发语言的空间分析建模法；②基于 GIS 外部松散耦合式的空间分析建模法；③混合型的空间分析建模法；④插件技术的空间分析建模法；⑤基于面向目标的图形语言建模法(图解建模)。

4. **图解建模法**：属于基于面向目标的图形语言建模法，用直观的图形语言将一个具体的过程模型表达出来。分别定义不同的图形代表输入数据、输出数据、空间处理工具，以流程图的形式进行组合并且可以执行空间分析操作功能。

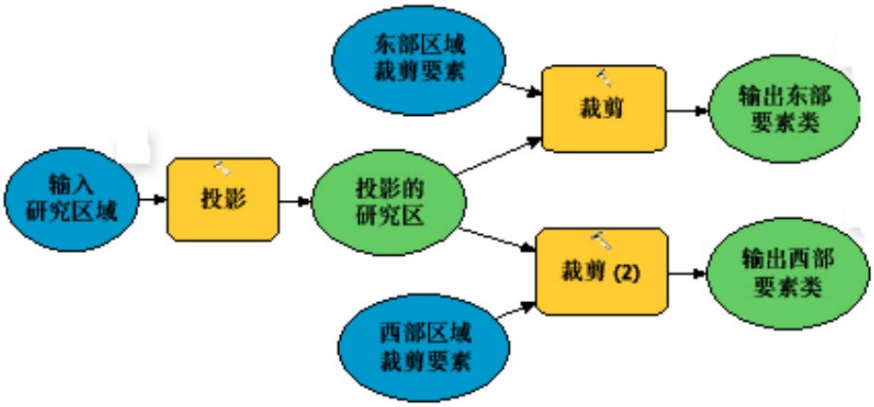
   - **ArcGIS 的模型生成器(Model Builder)的模型元素**

     1) 工具：地理处理工具是模型中工作流的基本组成部分。工具用于对地理数据

或表格数据执行多种操作。工具被添加到模型中后，即成为模型元素。

2) 变量：变量是模型中用于保存值或对磁盘数据的引用的元素。

3) 连接符：连接符用于将数据和值连接到工具。连接符箭头显示执行处理的方向。



- **模型生成器(Model Builder)的中间数据：**运行模型时，将在模型中创建各个流程的输出数据。某些输出数据只作为中间步骤创建，它们将连接到其他流程，并协助完成最终输出的创建。由这些中间步骤生成的数据称为中间数据。
- **模型参数：**模型参数是模型工具对话框中显示的参数。

第十三章·智能空间分析与空间决策支持系统

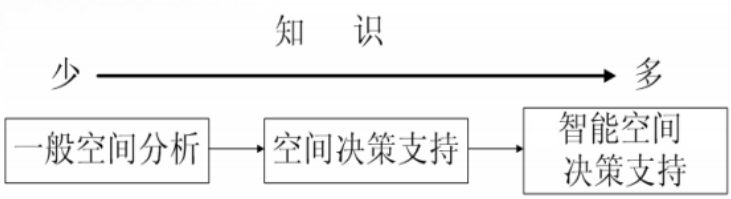
➢ 智能空间分析

1. **智能空间分析 = 人工智能 AI + 空间分析**。
2. **人工智能**：是智能机器所执行的通常与人类智能有关的智能行为，如判断、推理、证明、识别、感知、理 解、通信、设计、思考、规划、学习和问题求解等思维活动。
3. **智能空间信息处理(Intelligent Spatial Information Processing，ISIP)**：是指利用人工智能的理论和方法，利用计算智能方法，如神经计算、模糊计算、进化计算等方法实现空间信息的智能化处理，属于地球空间信息科学与人工智能的交叉与融合

➢ 空间决策支持系统

1. 根据空间分析的智能化程度和过程中引入知识的多少，空间分析划分为三种类型：

①一般空间分析 ➔ ②空间决策支持 ➔ ③智能空间决策支持



2. **空间决策支持的一般过程**：

   1) 确定目标：根据用户的要求，确定用户的最终实现目标，并对目标性质进行分类，确定目标的初步认识。

   2) 建立模型：建立分析的运作模型（用户实际运作过程中的各种业务运作模型）及定量模型（参照用户的实际工作模型，结合空间数据的空间特点，形成各种定量分析模型）

   3) 寻求空间分析手段：结合以上分析结果，逐步分解细节，寻求空间分析手段，对各种可能的分析手段进行分析，确定可行性的分析过程。

   4) 结果评价：对空间分析的结果进行评价，确定结果的合理性和可靠性。

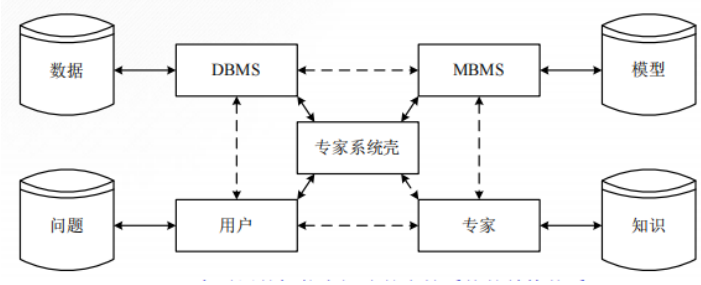
3. **空间决策支持与一般空间分析的区别：**①空间决策支持应用了多种分析运作模型和分析定量模型；②空间决策支持比一般空间分析具有更多的智能处理功能。

4. **空间决策支持系统（Spatial Decision Support System，SDSS）**

5. **智能空间决策支持：**在空间决策支持的基础上，增加了更多的人工智能技术，提高了空间决策支持的智能化处理水平，能够解决更加复杂的空间决策问题。

6. **通用的智能空间决策支持系统的结构体系**

使用软件工程和知识工程开发空间决策支持系统的开发环境(外壳或产生器)是建立空间决策支持系统的经济和灵活的方式。



1) **专家系统壳(Shell)：** 该系统的核心，可以单独作为专家系统开发工具，直接控制着 SDSS(空间决策支持系统)的控制流和对外交流的元知识，以及非结构化空间知识的推理机，是 SDSS 的大脑。为了便于使用空间数据和非空间数据，专家系统壳有一个与外部数据库的接口，包括 GIS，关系数据库和遥感信息系统。

2) **模型库管理系统：**管理和处理程式化知识，包括算法、统计程序和数学模型，有一个与专家系统壳的接口，可以通过专家系统壳的元知识进行调用。

7. **决策的定义：**决策是一个决策者为达到特定的目的，在一定的约束条件下，选择最优方案的过程

8. **决策问题的构成**：

1) **方案集合：**可供选择的决策方案集合，记为 A。

2) **状态集合：**决策问题所处的外界环境，称之为状态。系统所有可能的状态，称为状态集合，记为 Q。

3) **损益函数**：若采用策略$a(a \in A)$，系统状态出现$q(q \in Q)$，系统收益$W = (a, q)$定义映射：$W:(A \times Q) \to R$为决策问题的损益函数

损益表(*A*和*Q*可数)

	Q_1	Q_2	…	Q_n
A_1	W_{11}	W_{12}	…	W_{1n}
…	…	…	…	…
A_m	W_{m1}	W_{m2}	…	W_{mn}

4) **目标函数(决策准则)**：损益函数是系统的实际收益情况。给出收益的评价标准，即“抉择”时的优化准则。抉择准则对于不同的决策者、问题、方法都是不同的，它最终决定了方案的形成

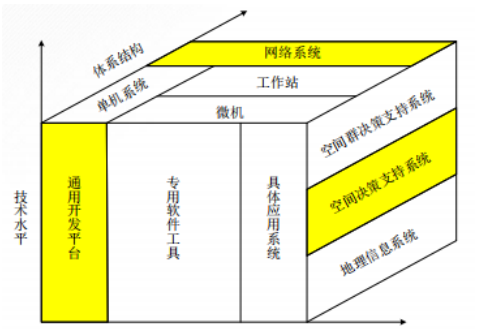
9. **决策问题的分类：**根据决策问题中 Q 的状态数，划分为以下类型：

1) 当系统状态集 Q 中状态数 n=1 时，为确定性决策问题；

2) 当 n>1 时，且系统各状态出现的概率未知时，为不确定性决策问题；

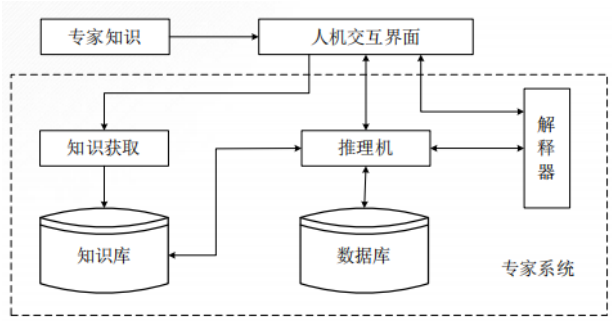
3) 当 n>1 且系统各状态出现的概率服从一个已知的概率分布时，为风险性决策问题

10. 根据空间决策支持系统的功能特点、技术水平和体系结构，研制的 SDSSP 定位在图中空间决策支持系统、通用开发平台和网络系统 3 个侧面相交构成的小立方体上，代表了当前的**空间决策支持系统的主要模式**。



➢ 空间决策支持系统的相关技术

1. **专家系统技术：**包括：数据库、知识库、推理机，解释器及知识获取 5 个部分。



2. **数据挖掘与知识发现：**知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程；数据挖掘是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤。

3. **空间数据挖掘(Spatial Data Mining, SDM)**：从空间数据中提取隐含其中的、事先未知的、潜在有用的、最终可理解的空间或非空间的一般知识规则的过程。➔ 李德仁：空间数据挖掘和知识发现(Spatial Data Mining and Knowledge Dicvoery, SDMKD)，能够把 GIS 有限的数据变成无限的知识，精练和更新 GIS 数据，促使 GIS 成为智能化的信息系统